Transportation Research Part A 135 (2020) 38–62

Contents lists available at ScienceDirect

Transportation Research Part A

journal homepage: www.elsevier.com/locate/tra

Quantifying the impact of urban road networks on the efficiency of local trips

Daniel Merchán*,[1], Matthias Winkenbach, André Snoeck

Massachusetts Institute of Technology, Cambridge, MA 02139, United States

ARTICLE INFO

Keywords:
Circuity
distance approximation
street network analysis
last-mile logistics
urban freight

ABSTRACT

City-level circuity factors have been introduced to quantify and compare the directness of vehicular travel across different cities. While these city-level factors help to improve the quality of distance approximation functions for city-wide vehicle movements, more granular factors are needed to obtain accurate shortest path distance approximations for last-mile transportation systems that are typically characterized by local trips. More importantly, local circuity factors encode valuable information about the efficiency and complexity of the urban road network, which can be leveraged to inform policy and practice. In this paper, we quantify and analyze local network circuity leveraging contemporary traffic datasets. Using the city of São Paulo as our primary case study and a combination of supervised and un-supervised machine learning methods, we observe significant heterogeneities in local network circuity, explained by dimensional and topological properties of the road network. Locally, real trip distances are about twice as long as distances predicted by the L_1 norm. Results from São Paulo are compared to seven additional urban areas in Latin America and the United States. At a coarse-grained level of analysis, we observe similar correlations between road network properties and local circuity across these cities.

1. Introduction

Analytical approximation methods are widely used to quantify travel distances of vehicles within a transportation system. They can be applied to large-scale networks very efficiently, as their data requirements are typically limited to only a few parameters, such as basic geospatial information (i.e., latitude and longitude coordinates) of points of demand (PODs). Analytical distance approximations are particularly useful to inform decisions related to the strategic design and planning of transportation and logistics systems. In such decisions, the focus of the analysis lies less on an exact result for a specific realization of customers to be served, but more on the expected performance of the system.

In the design of urban transportation systems, the so-called L_1 or *rectilinear* norm is a common distance metric assumed when analytically approximating vehicular travel distances within the underlying road network. This norm assumes that the road network resembles a perfectly rectangular lattice. However, real-world urban road networks rarely exhibit consistent and perfectly rectangular designs. Several authors have thus shown that using the *Euclidean* or L_2 norm, conditioned on the proper estimation of a detour or *circuity* factor, as the distance metric in analytical approximations yields superior results, as it more appropriately accounts for properties of the underlying road network that affect travel directness and, consequently, route efficiency (Love and Morris, 1979).

* Corresponding author.
E-mail addresses: dmerchan@mit.edu (D. Merchán), mwinkenb@mit.edu (M. Winkenbach), asnoeck@mit.edu (A. Snoeck).
[1] This work was completed while the first author was a doctoral candidate at the Massachusetts Institute of Technology.

https://doi.org/10.1016/j.tra.2020.02.015
Received 7 December 2018; Received in revised form 9 February 2020; Accepted 19 February 2020

Love and Morris (1972, 1979) and more recently Ballou et al. (2002) and Giacomin and Levinson (2015) introduce circuity factor estimates for approximating country-level and city-level travel distances. In addition to improving the calibration of analytical distance approximations, circuity factors can be leveraged to assess the overall travel efficiency of the urban road network, and thus better inform design and planning choices about transportation and logistics systems operating on those networks.

Existing city-level circuity estimates are typically quantified based on commuter travel patterns, covering a wide range of possible inter-stop distances. For instance, Giacomin and Levinson (2015) study trips of up to 60 km. Nevertheless, the design and planning of urban transportation systems characterized by relatively short, localized trips demand more granular measurements of network circuity. Examples of such systems include large-scale (i.e., real-world) last-mile delivery systems, in which average inter-stop distances within a network can be as short as 0.2 to 0.3 km. Shorter trips tend to be more circuitous (Levinson and El-Geneidy, 2009) as the effect of road network obstacles (e.g., highways, rivers) and road network complications (e.g., one-way streets) on travel efficiency amplifies. Furthermore, cities generally exhibit significant differences in road infrastructure, in complications to travel directness and in urban form across zones. These differences are hardly characterized by a single city-level circuity factor. Thus, localized measures of road network circuity are needed. To the best of our knowledge, these granular measurements are not available in the existing literature yet. In addition, the correlation between circuity and road network properties at the local level remains largely unexplored. Previous studies (see, e.g., Ballou et al., 2002) have outlined features of the road network that affect circuity. Nonetheless, the extent and, consequently, relevance of these effects has not yet been quantified.

The lack of literature on local road network circuity may be due to the relatively recent interest in solving large-scale last-mile logistics problems, driven mostly by the continuous growth of cities and also by the rapid expansion of online retailing and, consequently, of last-mile delivery and pick-up services. In 2015 alone, parcel deliveries grew at 7–10% in mature markets and up to 300% in developing markets such as India (Joerss et al., 2016). The fast and dynamic evolution of e-retailing is driving structural changes in the way companies are reaching the urban consumer: as consumer expectations in terms of delivery speed and service options continue to diversify, online retailers are pressured to provide delivery service offerings that vary in several dimensions including product exchange location (e.g., deliveries to homes or to pick-up points) and transportation agents (e.g., parcel operators or crowd-sourced, ride-sharing services) (Winkenbach et al., 2018; Janjevic et al., 2019; Janjevic and Winkenbach, 2020).

Evidently, logistics flows in urban areas will continue to intensify, putting more pressure on an already congested urban road network. Estimates from developed markets predict that logistics flows already amount to approximately a quarter of vehicle miles traveled in a typical city and generate 16–50% of the pollutant emissions from transportation-related activities, depending on the type of pollutant considered (Dablanc, 2007). There are reasons to believe that these figures are higher in cities in emerging markets, given these regions' lower transport and logistics productivity levels and the fragmentation of retail channels. In particular, the prevalence of the *nanostore* channel, characterized by small, non-organized and family-operated retail formats, intensifies freight activities in already dense and congested urban areas as more frequent replenishments are needed due to store space or financial constraints (Fransoo et al., 2017; Blanco et al., 2017). Putting differences in urban freight operations between developed and emerging markets in perspective, a large beverage manufacturer and distributor serves nearly $7,000 - 8,000$ retailers on a weekly basis in New York City, and between $60,000 - 70,000$ retail establishments per week in São Paulo, Brazil. Thus, research efforts geared towards better informing the design and planning of last-mile distribution systems entail a significant societal impact.

The traditionally onerous effort to obtain reliable road network and traffic data has also limited the possibility to derive an in depth and quantitative understanding of the local efficiency of the urban road network. Nonetheless, large traffic and geospatial datasets extracted from contemporary mapping and navigation tools offer a window of opportunity to quantify and study the efficiency of urban road networks at the local level. This approach resonates with the vision of a science of cities, as proposed by Batty (2013), observing them through a complex system lens and leveraging new methods for data-driven studies of urban planning problems.

In this paper, we quantify and analyze the circuity of the urban road network for shortest path and minimum distance local trips. Further, building on a data-driven, network-theoretical approach and a combination of supervised and unsupervised machine learning methods, we analyze the topological and dimensional properties of the road network that affect local travel directness. Our analysis aims at deriving general correlations between the efficiency of the road network and its dimensional and topological properties. In doing so, we make the following contributions:

(i) We derive empirical estimates of road network circuity at a geographical scale and resolution that is relevant for last-mile logistics operations.
(ii) We propose a data-driven approach based on unsupervised machine learning models to classify urban areas according to their topological and dimensional properties.
(iii) We introduce a quadratic regression model to derive general correlations between the local circuity of the road network and its topological and dimensional properties.

The metropolitan area of São Paulo, Brazil, serves as the primary illustrative example for the methods presented in this paper. Results from São Paulo are compared and contrasted with other cities in Latin America and the United States (US).

We argue that an in-depth and quantitative understanding of the properties of the road network that affect circuity can inform logistics design and planning in several dimensions. Logistics practitioners can use more accurate, local circuity estimates to better approximate distances traveled in the road network and, consequently, better plan vehicle routes and fleet capacities. Furthermore, a better understanding of the complexity and efficiency of the road network should inform strategic decisions such as the design of delivery territories, the vehicle type choice, and the location of logistics facilities. The results of this study also render valuable

insights for policy makers as it explores the correlation between network efficiency and urban design decisions (e.g., defining the road network layout), or traffic management interventions (e.g., implementing one-way streets).

The remainder of this paper is structured as follows. In Section 2, we summarize the extant literature on analytical distance approximation methods, network circuity, and street network analysis. Section 3 introduces a transferable method to quantify network circuity at the local level using contemporary traffic datasets. In Section 4, we present a polynomial regression model to explore the impact of dimensional and topological features of the road network on network circuity. Section 5 explores the transferability and generalizability of our findings by comparing them across additional case studies. We conclude the paper with a discussion in Section 6.

2. Background

In this section, we review the extant literature on distance estimating functions and road network circuity. We focus our discussion on existing circuity estimates for urban travel. We also review recent studies on applications of network science to street network analysis.

2.1. Distance estimating functions

The following general form function has been widely used to approximate the distance between two points p, q in geographical space (Love and Morris, 1972):

$$d(p, q) = c\left[\sum_{i=1}^{2} |p_i - q_i|^r\right]^{\frac{1}{s}} \tag{1}$$

with parameters c, r and s. Parameter c quantifies the *circuity* of the underlying network, that is to say, the complications to travel directness. The circuity parameter c holds particular interest to this study and we formally define it in Section 2.2. Assuming $c = 1$, the Euclidean (L_2) and rectilinear (L_1) norms are special cases of this general form by setting $r = s = 2$ and $r = s = 1$, respectively.

Based on empirical results for inter-city distances, Love and Morris (1972) observe that setting $r = s$ provides the practical benefit of having to fit one less parameter at limited accuracy expense. Also, $r = s$ yields a convex function, which is a desirable property for computational purposes in a wide range of modeling applications, including facility location models. In a subsequent work, Love and Morris (1979) provide empirical evidence on the accuracy of this distance estimating function for intra-city travel. Their results suggest that, given a properly fitted value for c, the Euclidean norm generally outperforms the rectilinear norm also for urban travel distance estimations, unless the road network is consistently rectangular. A discussion on a weighted L_2-L_1 norm is provided in Brimberg and Love (1992).

Distance approximations have also been introduced in the context of routing problems. There is an extensive body of work on the use of continuum approximation (CA)-based models to approximate the expected distances of traveling salesman and vehicle routing problems for idealized network topologies (Beardwood et al., 1959; Daganzo, 1984a, 1984b; Newell and Daganzo, 1986a, 1986b). Several extensions to these models have been studied to account, for instance, for different area sizes and shapes, the number of customer locations, and the effect of time-windows (Chien, 1992; Kwon et al., 1995; Figliozzi, 2009). Building on CA-based models to approximate routing costs, Smilowitz and Daganzo (2007) present an optimization framework to design large-scale package distribution systems. Winkenbach et al. (2016a) further extent the use of routing cost approximations by introducing an augmented routing cost expression to account for maximum service time constraints within a mixed-integer linear programming model. This model is used to solve the capacitated two-echelon location-routing problem (2E-CLRP) for designing a large-scale urban logistics network (Winkenbach et al., 2016b). We refer the reader to a recent paper by Ansari et al. (2018) for a comprehensive overview on the evolution of CA-based methods applied to logistics and transportation systems modeling, including routing problems, over the past two decades. Nevertheless, the focus of these studies continues to be on the use of idealized road networks, mainly the L_2 and L_1 norms.

2.2. Network circuity

Circuity measures the relative detour incurred by vehicles traveling within a network compared to the straight-line distance between the origin and the destination of their path. A circuity factor c is thus defined as the ratio between the shortest-path network distance d_N and the Euclidean distance d_{L_2} such that

$$c = \frac{d_N(p, q)}{d_{L_2}(p, q)}, \tag{2}$$

for any pair of path origin and destination locations (p, q). This factor is equivalent to the inflation parameter introduced in Love and Morris (1972). A factor closer to 1.0 indicates higher levels of network efficiency (Barthélemy, 2011).

Theoretically, for intra-city distances, if travel is assumed to occur over an isotropic, rectilinear grid (i.e., a rectangular lattice), then the extant literature suggests $\bar{c} \approx 1.27$ (Larson and Odoni, 1981). Love and Morris (1979) empirically find values for c between 1.16 and 1.28 for selected urban areas in the US, and *circa* 1.35 for rural zones. Similarly, Newell (1980) estimates a factor of $\bar{c} = 1.20$ for general urban travel.

Levinson and El-Geneidy (2009) use c to analyze the selection of residential locations for commuters. In their study of 22 cities in the US, they find an average $\bar{c} = 1.18$. Using the Minneapolis - Saint Paul region for an in depth study, they report a circuity factor of 1.58 for travel distances less than or equal to 5 km. Through regression analysis, they explain city-level road network circuity based upon a set of network attributes, such as the number of street-to-street and freeway-to-freeway nodes, street length, and freeway length for a 2 km buffer around the line representing the Euclidean distance of a trip. Model results suggest that street and freeway length decrease circuity, i.e., the larger the road length, the higher the likelihood of a direct trip between origin and destination. On the contrary, they observe that the number of street-to-street and freeway-to-freeway nodes increase circuity. This is expected as in highly dense zones (i.e., large number of nodes), trips are more circuitous. However, the low $R^2 = 0.11$ of the model limits its explanatory power.

Giacomin and Levinson (2015) empirically estimate $\bar{c} = 1.34$ for the 51 most populous metropolitan areas in the United States and find statistically significant evidence of road network efficiency decline between 1990 and 2010 for nearly 70% of the metropolitan areas. Circuity estimates are weighted by distance traveled in home-to-work commutes considering trips of up to 60 km, based on the US National Travel Household survey (United States Department of Transportation, 2009). As expected, they also observe that circuity decreases inversely proportional to distance, which is also concluded in Levinson and El-Geneidy (2009). Using the city of Stuttgart, Germany, as their case study, Ehmke and Campbell (2014) suggest a factor of 1.5 to correct straight-line distance estimates between downtown and suburban areas to inform order-acceptance mechanism for home-delivery services but provide not further references on how this value is derived. Huang and Levinson (2015) use circuity to investigate transportation mode choice for commuters and observe that transit networks, which prioritize spatial coverage at the expense of directness, usually exhibit higher levels of circuity compared to road networks.

Network circuity has also been explored for inter-city travel. Love and Morris (1972) observe values between 1.16 and 1.18 in the US. Ballou et al. (2002) analyze inter-city circuity in different countries. They find that c ranges between 1.12 and 2.10, depending upon road density, connectivity and geographic obstacles, but provide no further analysis on the relative importance of each of these factors. We summarize existing relevant circuity estimates in Table 1, also noting that the majority of studies have focused on cities in the continental US.

Merchán and Winkenbach (2019) propose a data-driven extension to calibrate the CA-based models to better approximate route distances introduced by Daganzo (1984b) based on empirically derived local circuity factors using real-world traffic datasets. They conclude that the circuity of the underlying road network has a significant impact on the predictive performance of CA-based methods in real-world urban settings. Building on their work, Bergmann et al. (2020) use regression analysis to determine closed-form correction factors for CA-based models to account for the impact of integrating pickups and deliveries on joint urban vehicle routes.

2.3. Street network analysis

Network (graph) theory is a widely used lens to approach the analysis of urban street networks. In fact, its use dates back nearly three centuries with Euler's classic seven-bridge problem at Königsberg (now Kaliningrad) (Barabási, 2016). Fundamentally, a network is a finite set of *nodes* (or *vertices*), connected by a finite set of *links* (or *edges*). The orientation of the links determines if the network is directed, undirected, or mixed. In urban transportation networks, links commonly represent streets and nodes represent street intersections and *cul-de-sacs*. This representation is usually know as *primal* (Porta et al., 2006). Alternatively, the dual approach models streets as nodes and intersections as links. Even though the primal provides a more intuitive representation of the street network, the dual representation is at the core of the popular *space syntax* method first introduced by Hillier and Hanson (1984) and has been used in subsequent works, such as in Jiang and Claramunt (2004). A comparative analysis of both representations is available in Porta et al. (2006) and Porta et al. (2006).

A *spatial network* is a network embedded in a (usually two or three) dimensional space and characterized by a metric (usually the Euclidean distance). This distinction is relevant as the spatial constraint on networks has relevant implications on its topological and dimensional properties. The urban road network is usually modeled as a spatial and approximately planar network (Barthélemy, 2011).

Advances in geographic information systems and new sources of data are triggering new frontiers of quantitative analysis of urban

Table 1
Survey of circuity factor estimates in the extant literature.

c	Geographical Scale	Case study	Source
1.27	Urban	–	Larson and Odoni (1981)
1.16–1.28	Urban	US	Love and Morris (1979)
1.35	Rural	US	Love and Morris (1979)
1.20	Urban	US	Newell (1980)
1.18–1.58	Urban	US	Levinson and El-Geneidy (2009)
1.34	Urban	US	Giacomin and Levinson (2015)
1.50	Sub-urban	Germany	Ehmke and Campbell (2014)
1.16–1.18	Country	US	Love and Morris (1972)
1.12–2.10	Country	Worldwide	Ballou et al. (2002)

infrastructure (Batty, 2013). In particular, there has been an increasing interest in the literature to approach the study of urban road networks as *complex spatial networks* and analyze them from a large-scale quantitative standpoint (see, e.g., Barthélemy, 2011, and references therein). For instance, in spite of the very different and varied processes shaping cities, unexpected quantitative similarities have been found at least at the coarse-grained level (Jiang and Claramunt, 2004; Crucitti et al., 2006; Lämmer et al., 2006; Barthélemy and Flammini, 2008; Louf and Barthelemy, 2014).

A complex network is described by a set of topological measures that characterize its structure, i.e., its connectivity, centrality, and resilience. Two commonly used *connectivity* measures include *node degree* and *node connectivity*. *Node degree* measures the number of edges (i.e., streets) incident to a node. Due to planar constraints, urban street networks exhibit low variability in node degree measurements, ranging usually between 2 and 4 (Lämmer et al., 2006). The *node connectivity* of a network measures the minimum number of nodes that must be removed to disconnect the graph. In street network analysis this measure is frequently equal to 1 due to the presence of *cul-de-sacs*. Thus, a more useful alternative is to use the *average node connectivity*, which measures the expected number of nodes that must be removed to disconnect a random pair of non-adjacent nodes (Boeing, 2017).

Centrality measures inform the importance of nodes, and consequently, the resilience of a network. For instance, *betweenness centrality* for a node j is measured as the ratio of the number shortest paths going from node s to node t passing through node j, over the total number of shortest paths going from s to node t. The spatial distribution of the betweenness centrality encodes relevant structural information and can be used to quantify the suceptibility of the network to traffic congestion (Barthélemy, 2011). Other centrality measures include *closeness* and *degree centrality*. By using centrality measures, Porta et al. (2009) observe the relationship between zones with better centrality and the location of commercial establishments in Bologna.

Connectivity and centrality measures characterize the topology of the urban street network. Nevertheless, given the highly heterogeneous geometries of street networks, a purely topological perspective is insufficient to fully characterize a street network (Louf and Barthelemy, 2014). As noted by Ratti (2004), a richer understanding of the urban texture arises when the still-valid simplifications of the space syntax framework from a topological perspective are combined with dimensional analysis (see Fig. 1). Dimensional measures inform the spatial distribution of nodes and include intersection density, edgedensity, street length, diameter and circuity. We refer the reader to the work by Barthélemy (2011) for a comprehensive overview of spatial networks and their application to transportation and infrastructure systems, and to the manuscript by Boeing (2017) for a comprehensive overview of topological and dimensional measures.

2.4. Literature gap

Previous studies in road network circuity have focused either on inter-city trips or intra-city commuter trips (see Table 1). Nevertheless, the nature of large-scale last-mile logistics, characterized by short-distance trips, demands more granular circuity measurements. Local trips tend to be more circuitous as the effect on travel efficiency of road network obstacles (e.g., highways, rivers) and road network complications (e.g., one-way streets) is more profound. Furthermore, cities generally exhibit significant differences in topology, infrastructure, obstacles and complications to travel directness across their various neighborhoods or zones, which can hardly be characterized by a unique, city-level circuity estimate. Giacomin and Levinson (2015) also suggest that future studies should address the causal relations of network circuity. This study targets both of these gaps in the extant literature.

3. Quantifying local road network circuity

In this section, we first outline a data-driven approach to delimit the urban area of interest based on population density measurements, and define the unit of geo-spatial analysis used to segment the urban area. Second, we describe the sampling method to quantify road network circuity for local trips using real road network datasets. We conclude this section with a discussion of the

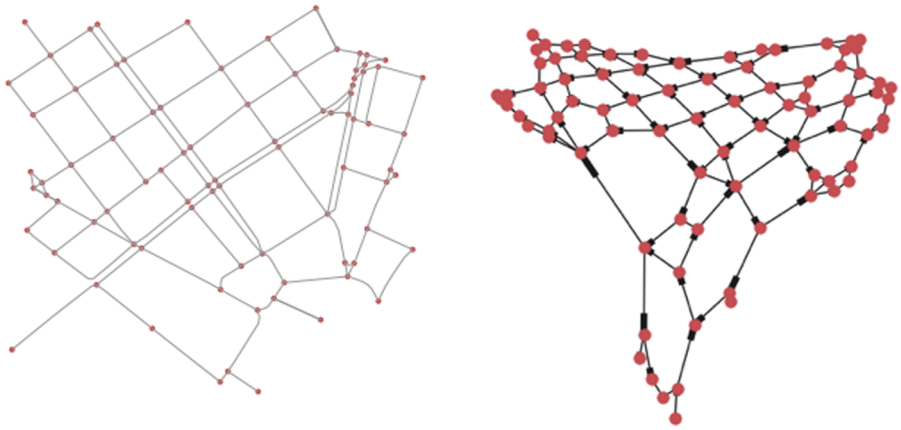


Fig. 1. Dimensional and topological representations of a segment of the road network. While dimensional measures characterize the spatial distribution of the network, topological measures define its connectivity and structure.

results of the sampling methods applied to our case study.

3.1. Unit of geo-spatial analysis

Urban areas of interest usually extend beyond official city boundaries, requiring a certain degree of arbitrariness to define them. Consider, for instance, the case of São Paulo. On the one hand, if we limit our attention to the municipal boundaries, many relevant and densely populated surrounding zones will be excluded. Urban population polycentricity is a common characteristic of large metropolitan areas. On the other hand, if we consider the entire Metropolitan Region of São Paulo, it covers an area of nearly 7,000 km^2, including numerous low-density zones, which are of scant interest to our analysis. To find a middle ground, we use a population density threshold to discriminate areas of interest. Even though this approach is arbitrary to some extent, it also easily scalable and transferable, given a reliable and consistent source of population data. *LandScan*, a global population database developed by the Oak Ridge National Laboratory (Bright et al., 2015) based on high-resolution satellite imagery, is our source of universally available population data. It provides up-to-date ambient population counts at a spatial resolution of approximately 1 square kilometer (km^2). We build our analysis on data from the 2015 *LandScan* database.

The urban area of study is divided into an grid of square segments to discretize our geo-spatial data and analysis. This simple segmentation and data aggregation approach, also known as raster data model (Singleton et al., 2018), is appealing as it facilitates intra-city and inter-city comparisons, independently of any local administrative divisions (e.g., zip-codes or cadastral zoning). The choice of segment size needs to balance data resolution with data processing efficiency, which varies across applications.

3.2. Trip distance calculation

We consider minimum distance paths obtained form the Google Distance Matrix (GDM) web service (Google, 2017). Temporal dependencies such as congestion or customer time-windows, and alternative objective functions (see, e.g., Figliozzi, 2008) which may impact local circuity, fall outside the scope of our analysis. To the best of our knowledge, the specific shortest-path algorithm supporting the GDM web service has not been officially disclosed by Google. Bast et al. (2016) report *Transfer Patterns* (Bast et al., 2010) as one of the algorithm used for public transportation routing in Google's products. This efficient technique particularly for multi-modal trips breaks down the problem into transfer patterns (i.e., sequences of stops where transportation mode changes occur) and then uses Dijkstra's algorithm (Dijkstra, 1959) or other efficient methods to find the shortest-path for single-mode direct connections. We refer the reader to the manuscript by Bast et al. (2016) for a comprehensive survey of shortest-path algorithms in road-networks, including but not limited to goal-directed methods, hierarchical techniques and labeling algorithms.

3.3. Sampling and circuity factor estimation

Within each square segment i, we generate T_i random and uniformly distributed origin-destination points, snap them to the nearest street segment, and obtain the point-to-point *shortest-path* trip distances form the GDM. We define T_i based on the sampling method described in Law and Kelton (2000) to estimate average values given a specified absolute error ϵ. Specifically, T_i is the minimum sample size for which the *t-test* confidence interval half-length with a confidence level α is less or equal than ϵ. Next, we obtain c_{it} for each t trip using Eq. (2). Finally, we quantify the circuity factor for each segment c_i using the following expression:

$$c_i = \sum_{t=1}^{T_i} c_{it}/T_i. \tag{3}$$

We emphasize the value of the GDM service for transportation and urban planning research. While the use of geographic information system (GIS) tools to estimate travel distances by researches and practitioners is not novel, the use of classic GIS tools has been constrained by the usually limited availability of reliable cartographic information, particularly in emerging markets. Contemporary distance and traffic data sources such as GDM, and geo-spatial data sources such as OpenStreetMaps (OSM) (The OpenStreetMap Foundation, 2017), discussed in detail in Section 4, offer under-explored opportunities to efficiently collect and process worldwide and up-to-date urban road infrastructure and traffic information, enabling scalability and transferability of methods.

Finally, we note that our sampling approach to estimate circuity differs from the method described in Boeing (2017), which 'relocates' sampled origins and destinations points to the nearest node (i.e., road intersections) in the network. As expected, this method biases circuity estimates as road intersections tend to be more accessible than any other random points within a road segment. We argue that our sampling approach therefore better represents the real-word circuity properties of short, local trips.

3.4. Application

The core of São Paulo's metropolitan area, including the municipality of São Paulo and its surroundings, serves as our primary illustrative example. As noted in Section 3.1, to focus our analysis on the most relevant zones within a metropolitan area, we select urban segments with ambient population density of at least 1000 inhabitants/km^2. We derive this population density threshold based on preliminary data exploration. The resulting urban area covers approximately 1630 km^2 and encompasses approximately 85% of the 20 million inhabitants within the metropolitan area. Furthermore, we choose city segments to have a size of 1 km^2 each, to ensure sufficiently detailed spatial resolution and consistency with the population data source.

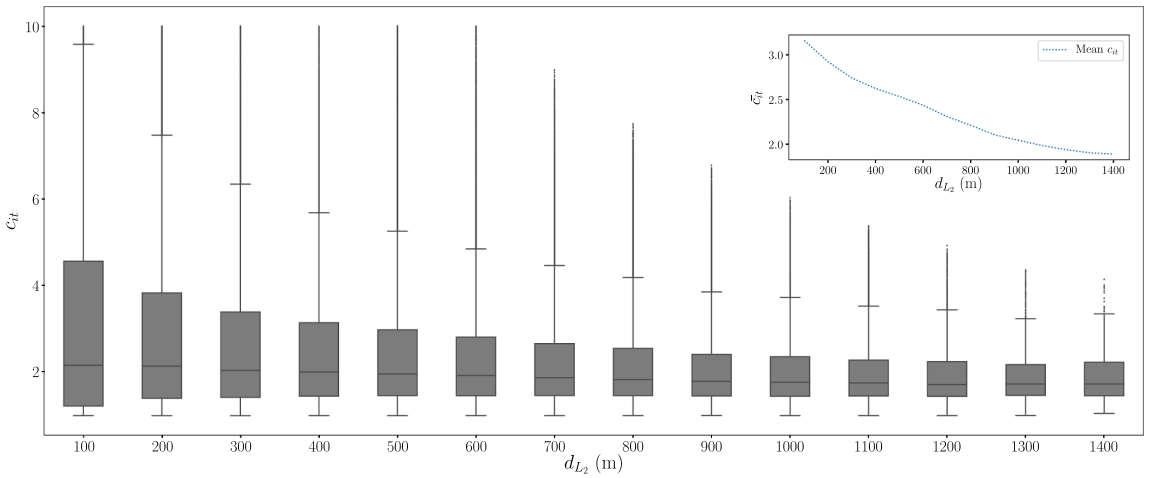


Fig. 2. Negative, non-linear and asymptotic relationship between trip-level circuity, c_{it}, and trip Euclidean distance d_{L_2}.

We simulate and process $T \approx 190$ ($\epsilon \leqslant 0.15$) local trips per segment according to the above mentioned sampling method and obtain c_{it} for each trip t and each segment i per Eqs. (2) and (3), respectively. In total, we process approximately 312,000 trips. The average real network trip distance is 1.16 km, the median distance is 0.99 km and the upper bound is approximately 5 km. As a result of the area size of 1 km^2 defined for each segment, the Euclidean distance of each trip, d_{L_2}, is bounded at $\sqrt{2}$ km. At the trip-level, we observe a negative, non-linear and asymptotic relationship between c_{it} and d_{L_2} (see Fig. 2). The corresponding variability reduces in d_{L_2}, suggesting a more profound and less predictable effect of road network complications on shorter trips.

The segment-level circuity factor c_i ranges from approximately 1.38 to 5.32 with an average of 2.51, a median of 2.34 and an inter-quartile range of 0.90 (see Table 2), which indicate a positively skewed distribution of c (see Fig. 3). Based on a Kolmogorov-Smirnov (KS) goodness of fit test (p-value = 0.28), the distribution of c fits a *lognormal* distribution with parameters 0.51, 1.15 and 1.20 for shape, location and scale, respectively. This result suggests that the average local circuity factor based on real trips, $\bar{c}$, is nearly twice as large as the analytically derived factor of 1.273 assuming travel according to the L_1 metric (Larson and Odoni, 1981), and significantly larger than those reported for city-level trips (see Table 1).

We observe particularly high circuity levels towards the inner parts of the city, in zones crossed by major road obstacles, such as highways, and in peripheral segments (see Fig. 4). Higher levels of circuity in peripheral segments are to be expected as these areas usually exhibit network topologies that resemble tree-like structures instead of well-connected road grids. In inner city zones, in spite of having higher levels of network connectivity, also exhibit higher levels of circuity due to one-way streets and other complications to travel. The relationships between circuity and other road network properties are explored in detail in Section 4. Interestingly, in these same inner city segments with less efficient road networks, the intensity of local trips is usually larger as a result of higher levels of ambient population density. That is to say, a large portion of local trips (e.g., local deliveries in logistics operations) take place in city areas with highly circuitous road network infrastructure.

4. Explaining local road network circuity

In this section, we explore properties of the urban road network that impact network circuity. Using the primal representation of street networks, we define a set of dimensional and topological variables to characterize the road network of any city segment and analyze them as explanatory variables of the circuity factor c. A cluster analysis based on a Gaussian mixture model (GMM) serves as a starting point to generate a classification of segments. We then introduce a quadratic regression model to explore the relationship between the explanatory variables and c. Considering again the São Paulo example, we use the estimates of c_i for each city segment i derived in Section 3.4 as values for the dependent variable. Measurements for the potential explanatory variables are obtained from OSM according to the data processing method described in Section 4.1 below.

4.1. Potential explanatory variables

As discussed in Section 2, dimensional (i.e., metric) variables describe physical properties of the road network, whereas

Table 2
Summary statistics for c across São Paulo.

Mean ($\bar{c}$)	Coeff. Var.	Median	Quartile 1	Quartile 3	Inter-quartile range
2.51	0.28	2.34	2.00	2.90	0.90

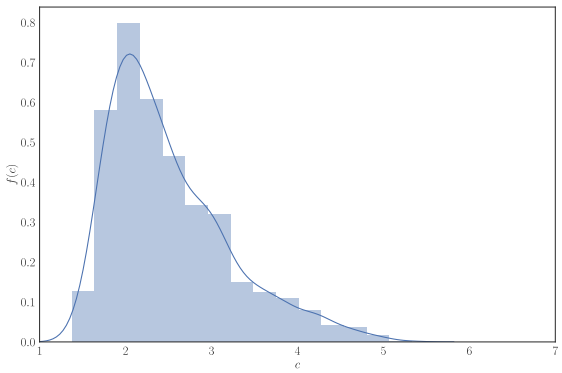


Fig. 3. Positively-skewed distributions of c for the São Paulo dataset, based on 1630 segment-level measurements.

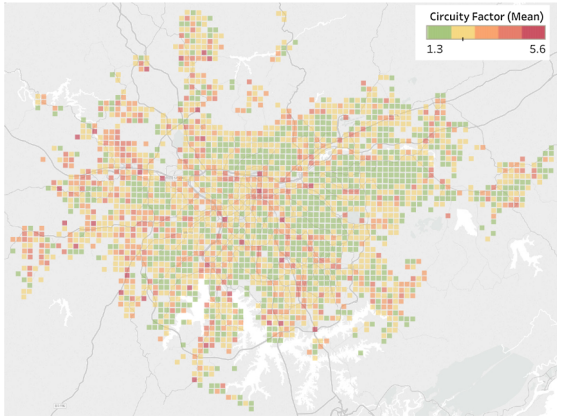


Fig. 4. Significant heterogeneities in network circuity observed across São Paulo, c ranges from 1.38 to 5.32, with an averge of 2.51.

topological variables characterize network connectivity, centrality and complexity. We argue that these physical and topological properties are correlated with the level of circuity of a given segment. Thus, building on choices of variables available in the extant literature (see Section 2.3), we define a set of dimensional (see Table 3) and topological variables (see Table 4) as potential explanatory variables of road network circuity. Formulae for topological variables are provided in Appendix B.

OpenStreetMaps (The OpenStreetMap Foundation, 2017) is the primary source of road network data. To process OSM data, we leverage the Python OSMnx module (Boeing, 2017). Three road types are defined in this study based upon their accessibility and traffic carrying capacity: highways, primary roads, and streets (see Table 3). Highways (highlighted in red and brick-red in Fig. 5) constitute the road type with the largest traffic carrying capacity, having at least 2 lanes in each direction, with some degree of separation and limited access. Primary roads (highlighted in orange in Fig. 5) represent the next most important road type, having usually 2–3 lanes in each direction and minimal or no separation. Major urban avenues are usually classified as primary roads in OSM. The third type, streets, groups the remaining road types for vehicle circulation in a city (highlighted in yellow and white in Fig. 5). These roads are characterized by no more than two lanes and are easily accessible, which facilitates travel directness.

Table 3
Segment-level metric variables.

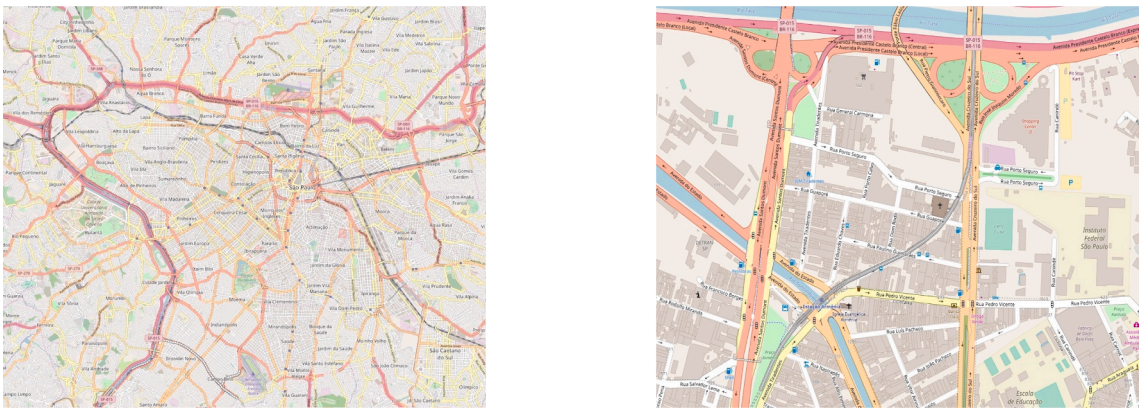
Variables	Description
Intersection density (/km^2)	Number of road intersections
Highway length (km)	Total length of highway roads
Primary road length (km)	Total length of primary roads
Street length (km)	Total length of non-highway and non-primary roads
One-way fraction (%)	Fraction of total street length with directional constraint (i.e., one-way streets)
Avg. road-link length (km)	Mean road-link length, including streets, primary roads and highways

Definitions adapted from Boeing (2017).

Table 4
Segment-level topological variables.

Variables	Description
Node connectivity	Average number of nodes to remove to disconnect a non-adjacent pair of random nodes
Node degree	Number of edges (streets) emanating from each node, averaged over all nodes
Neighborhood degree	Average node degree of a node's neighbors, averaged over all nodes
Betweenness centrality	Number of shortest paths that pass through a node, averaged over all nodes
Closeness centrality	Reciprocal of the sum of the distance from the node to all other nodes, averaged over all nodes
Degree centrality	Fractions of nodes that each node is connected to, averaged over all nodes

Definitions adapted from Boeing (2017).



(a) City-level (b) Zone-level

Fig. 5. City-level and zone-level road network extracts to illustrate the classification of roads: highways (red and brick red), primary roads (orange) and streets (yellow and white). Notice the significant presence of highways and primary roads within the city core of São Paulo. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

4.2. Classification of urban segments by means of cluster analysis

Given the large diversity of types of city zones in terms of dimensional and topological properties, we first conduct a cluster analysis, which is helpful to gain insights about the underlying structure of the data and to detect salient features (Jain, 2010). In this particular case, we leverage the cluster analysis to: (1) generate classes of city segments sharing similar road network characteristics, and (2) identify potential outliers, i.e., city segments with atypical road network properties, which could introduce significant bias to the analysis. Atypical segments include, for instance, zones with a scant road network coverage.

To generate clusters, i.e., archetypes of segments based on road network properties, we use a Gaussian mixture model with K-mixture components fitted using an expectation-maximization (EM) algorithm (Hastie et al., 2009). We select GMM as our clustering framework over its deterministic counterpart, K-means, since the non-deterministic assignment of observations to clusters in GMM using posterior probabilities offers additional information on the likelihood of each observation to belong to any of the K classes (Hastie et al., 2009). The GMM-based cluster analysis we conduct includes all metric and topological explanatory variables (see Tables 3 and 4), but does not include c. We expect this classification to inform preliminary correlations between road network properties driving the configuration of clusters and the circuity of the segments within those clusters.

In a pre-processing stage, we conduct a principal component analysis (PCA) on the explanatory variables to reduce the dimensionality of the dataset and address multi-collinearity issues among the explanatory variables. Further, the PCA provides useful information on the explanatory variables that account for the largest portion of the variance in the data, signaling which of these explanatory variables are most relevant. The number of principal components (PCs) to use for clustering is defined based on an explained variance threshold of $\phi = 0.9$ to balance model parsimony and explanatory power. We implement the PCA and the GMM in Python using the Scikit-learn module (Pedregosa et al., 2012).

The PCA yields preliminary insights about the underlying structure of the data. Out of the 12 initial explanatory variables, the first six PCs explain 93% of the variance in the data ($\phi = 0.9$). In analyzing the contribution of each explanatory variable onto the PCs (see Fig. 6), we make the following additional observations. Betweenness centrality, degree centrality, and connectivity-related variables (intersection density and street length) are the largest contributors to the first PC, explaining 41% of the variance. While connectivity and centrality related measures dominate the first PC (PC1), dimensional variables (i.e., complications to travel) are the largest contributor to the second PC (PC2) and explain 25% of the variance.

Based on the results of the PCA, we fit the GMM with $K = 3$ clusters (mixtures). We determine the value of K based on a cluster separation analysis using the silhouette score (Fig. 7). The largest cluster separation (i.e., highest score) is obtained by setting $K = 3$.

The spatial distribution of the resulting clusters is depicted in Fig. 8. We observe a first cluster, CL1 (red), composed mostly of

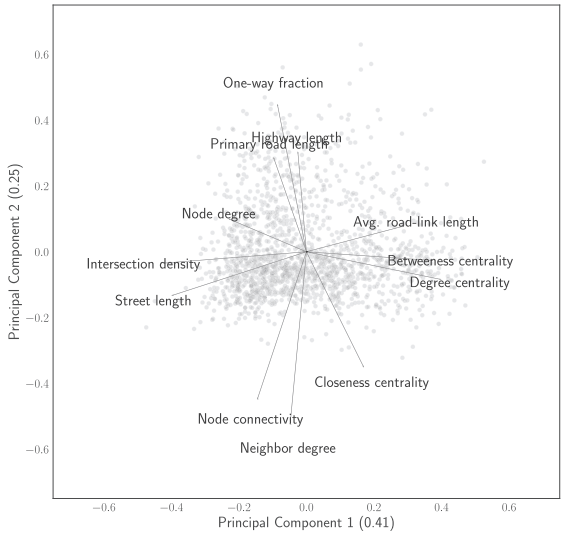


Fig. 6. The first two principal components account for 66% of the variance in the data. Variance in the first PC is mostly driven by centrality and connectivity variables, while variance in the second PC is mostly influenced by dimensional variables, i.e., one-way fraction and length of highway and primary roads.

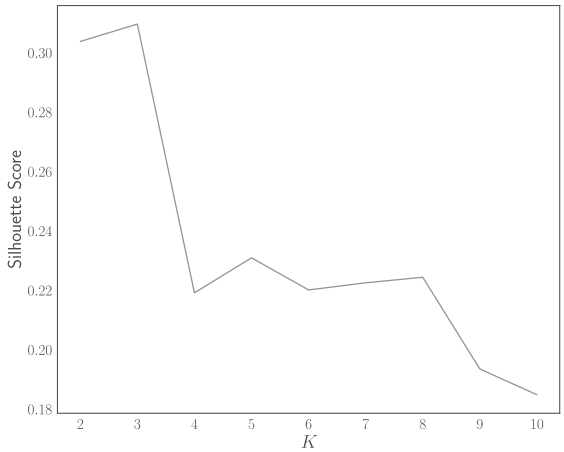


Fig. 7. Cluster separation analysis using silhouette score.

inner city segments and segments crossed by major highways and primary roads. A second cluster, CL2 (blue), is formed around outer city segments. The third cluster, CL3 (brown), corresponds primarily to peripheral zones and areas with limited or atypical road network infrastructure.

The spatial distribution of each cluster is compared against a projection onto the first two PCs (Fig. 9). Cluster CL1 corresponds to segments concentrated within the positive values of PC2 and negative values for PC 1, which, as observed in Fig. 6, correspond to segments exhibiting fine-grained road networks (higher node degree) with complications to travel (higher fraction of one-way streets and length of highway and primary roads). Thus, we refer to segments in this cluster as *constrained* road network segments. City segments corresponding to cluster CL2 are concentrated in the portion of the plot only driven by high network connectivity (PC1 <0). Therefore, we refer to segments corresponding to CL2 as *fine-grained* road network segments. Finally, segments corresponding to CL3 are concentrated within values PC1 >0, driven by higher network centrality, which typically resembles peripheral, less-developed areas. We refer to these segments as *coarse-grained* road network segments. Overall, the spatial distribution of the clusters (see Fig. 8) and the corresponding projections onto the main PCs (see Fig. 6) are consistent.

To further illustrate the distinction between clusters, Table 5 includes the average values per segment for a subset of explanatory variables. We obtain these average values considering all segments corresponding to a given cluster. Notice that we have also included in this summary c and three additional variables which were not used for clustering but provide additional information to compare clusters: fraction of urban area, fraction of population and mean population density.

The values reported in Table 5 yield preliminary insights on the correlation between explanatory variables and c. c is highest for CL1 and lowest for CL2. Segments corresponding to CL1 and CL2 have fine-grained road networks as indicated by the average node

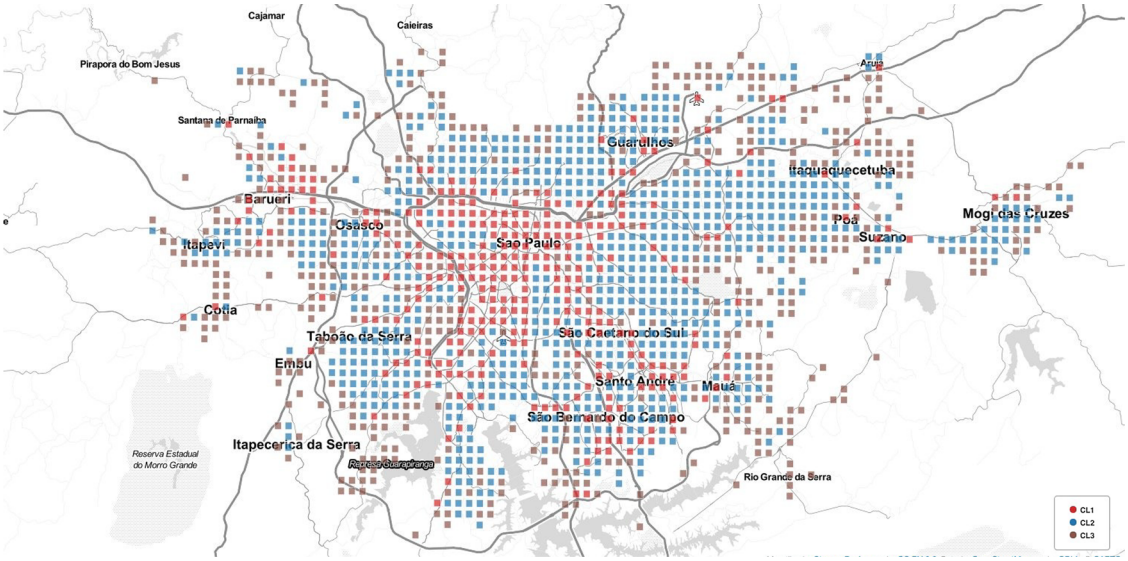


Fig. 8. GMM-based clustering results with $K = 3$ for the São Paulo data. A cluster (CL3) forms in the peripheries of the city grouping zones with atypical road network properties.

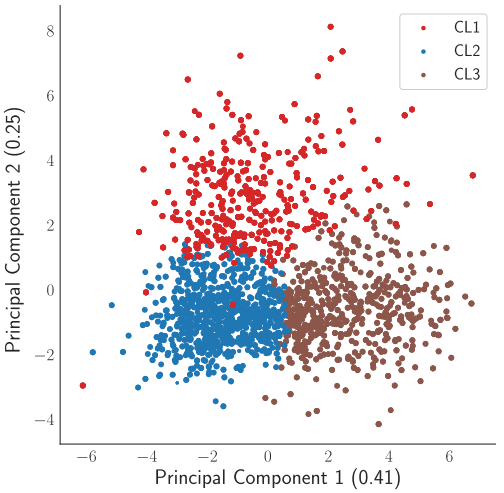
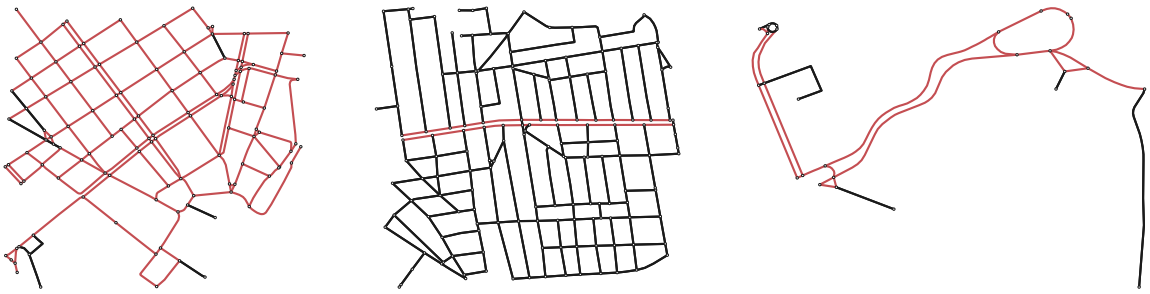


Fig. 9. Projection onto the main PCs of the segments classified by cluster type.

Table 5
Average values of segment-level variables for all three clusters.

		CL1	CL2	CL3
	Circuity factor $\bar{c}$	3.02	2.12	2.77
Explanatory variables	One-way fraction (%)	0.43	0.10	0.05
	Highway length (km)	964	90	80
	Primary road length (km)	1,518	362	228
	Node degree	3.10	3.03	2.82
	Node connectivity	0.97	1.57	1.32
	Betweenness centrality	0.07	0.06	0.11
Additional descriptors	Fraction of urban area	0.20	0.47	0.33
	Fraction of population	0.30	0.52	0.18
	Ambient population (inh/km^2)	15,005	11,550	6,146



(a) CL1 (b) CL2 (c) CL3

Fig. 10. Sample of road network segments in São Paulo for clusters CL1 (*constrained*), CL2 - (*fine-grained*) and CL3 (*coarse-grained*) with c = 2.80, 1.82, 4.49 respectively.

degree (3.10 and 3.03) respectively. Nonetheless, the mean node connectivity of CL2 is nearly 60% higher due to lower complications to travel compared to CL1, including but not limited to highway roads, primary roads and one-way streets. For instance, the average node connectivity of a segment with a bridge or an overpass will be low (even if the road network is fine-grained) as this feature will increase the probability of disconnecting the graph. Segments in CL3 are characterized by coarse-grained (lower node degree) and significantly more centralized networks compared to segments in the other clusters.

Finally, we select samples of typical road network configurations in city segments corresponding to each cluster to illustrate spatial differences in road network properties (see Fig. 10). The road network from CL1 (see Fig. 10a) and CL2 (see Fig. 10b) exhibit similar network connectedness. Nevertheless, circuity for CL1 (2.80) is nearly 50% higher due to directional constraints (red links) and the presence of highways and primary roads. The rightmost sample corresponds to CL3 (see Fig. 10c): its particularly high circuity factor (4.49) is driven mostly by its coarse-grained road network. Due to these atypical properties, segments corresponding to CL3 are excluded from the regression analysis presented in Section 4.3, which further explores the correlation between road network properties and circuity.

4.3. Regression analysis

Variable selection. The set of metric and topological variables (see Tables 3 and 4) exhibit strong correlations, which do not affect the clustering due to the use of PCA to de-correlate variables. In regression analysis, however, multicollinearity is undesired, as it inflates variances and, consequently, reduces the precision of coefficient estimates (Belsley et al., 1980). Several statistical test are combined to address multicollinearity among explanatory variables. First, we identify strongly correlated pairs of variables using the Pearson correlation coefficient (PCC) and select the key-covariates for the regression model. Further, we verify for multicollinearity in the regression analysis by means of two statistical tests: Variance Inflation Factor (VIF) and conditional indexes.

The variable selection step reduces the number of explanatory variables from 12 to 6 (see Table 6). For instance, average node connectivity is highly correlated with closeness centrality and street length (PCC of 0.76 and 0.60, respectively). The selection of key co-variates also prioritizes variables that are frequently used in the extant literature.

Overall, we observe non-linear correlations between the segment-level circuity factor, c, and each explanatory variable (see Fig. 11). As expected, circuity decreases as network connectivity in the corresponding segment increases. Circuity exhibits a positive correlation with the presence of obstacles and other complications to travel such as the fraction of one-way streets or the total length of primary roads and highways.

Regression model. To balance model complexity and interpretability of results, we introduce a polynomial regression model of second degree with interaction terms:

$$c = \beta_0 + \sum_{j=1}^{6} \beta_j X_j + \sum_{j=1}^{6} \beta_{6+j} X_j^2 + \beta_{13} X_1 X_2 + \cdots + \beta_{27} X_5 X_6 + \epsilon. \tag{4}$$

Standardized values are used given the significantly different measure scales that apply to each explanatory variable. We fit the regression model presented in Eq. (4) using the Python modules StatsModels (Perktold et al., 2017) and Scikit-learn (Pedregosa et al., 2012)

Table 6
Selected metric and topological most relevant co-variates for regression analysis.

Metric variables	Topological variables
X_1 Highway length (km)	X_4 Betweenness centrality
X_2 Primary road length (km)	X_5 Node connectivity
X_3 One-way fraction (%)	X_6 Node degree

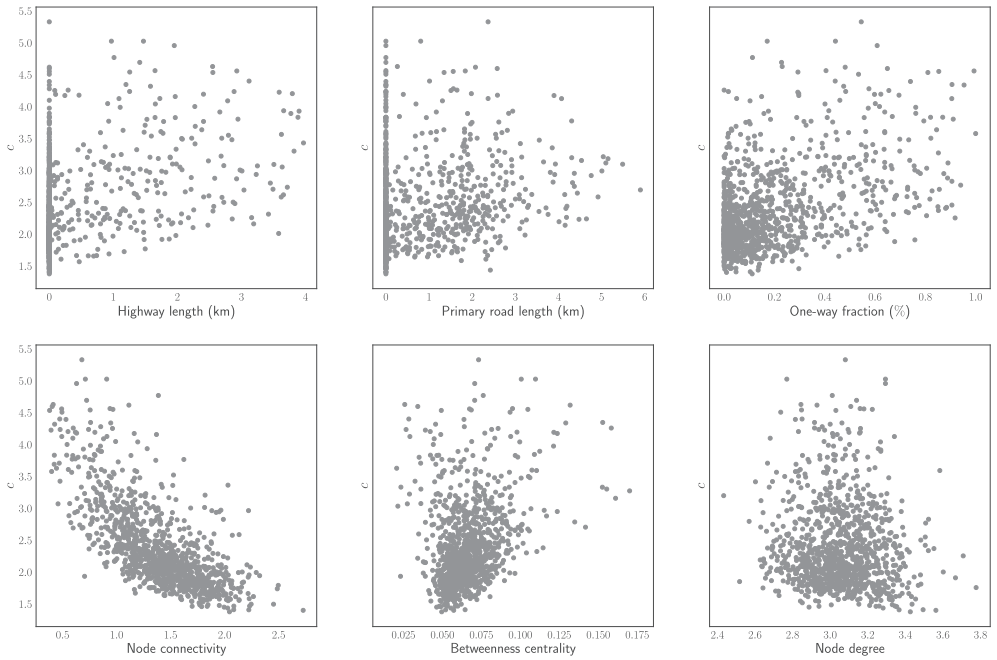


Fig. 11. Correlation between c and key explanatory variables for São Paulo, Brazil.

Table 7
Results form the regression model with standardized values. Statistically significant coefficients for $p < 0.01$. Column ΔR^2 reports the percent point (p.p.) difference in R^2 if the corresponding variable would be excluded from the model.

	Coeff. β	Std. Err.	p-value	ΔR^2 (p.p.)
Intercept	2.352	0.025	0.000	**
X_1 Highway length	0.225	0.031	0.000	−3
X_2 Primary road length	0.112	0.023	0.000	−2
X_3 One-way fraction	0.081	0.030	0.007	−2
X_4 Betweenness centrality	0.211	0.019	0.000	−7
X_5 Node connectivity	−0.274	0.026	0.000	−11
X_6 Node degree	−0.067	0.019	0.000	−2
X_1^2 Highway length × Highway length	−0.049	0.012	0.000	**
X_5^2 Node connectivity × Node connectivity	0.144	0.024	0.000	**
X_3X_4 One-way fraction × Betweenness centrality	0.040	0.018	0.005	**
R^2: 0.66	F-statistic:	74.15		
Adj. R^2: 0.65	Observations:	1077		

The results from our regression analysis (see Table 7) suggest that the presence of highways and primary roads exhibits the strongest positive correlation with the average circuity in a segment. This is expected, as at the local level, large-capacity roads usually complicate rather than facilitate travel directness. The magnitude of the standardized coefficient for highways is nearly twice as large the coefficient for primary roads as highways typically entails greater accessibility restrictions.

Our findings about the positive correlation between highway length and primary road length with circuity contrast with those reported by Levinson and El-Geneidy (2009), who find negative correlations. This difference evidences the necessary distinction between city-level and local circuity. For city-level trips, e.g., commuter travel and the 'line-haul' portion of a delivery route, large capacity roads facilitate travel directness (i.e., negative correlation with circuity). However, the opposite is true for local trips, e.g., the inter-stop portion of a delivery route.

The fraction of one-way streets further exhibits a positive correlation with local network circuity, which is also expected. Nevertheless, its magnitude is smaller compared to the effect of highways and primary roads. The interaction between one-way fraction and betweenness centrality is also significant: the effect on circuity of one-way streets amplifies for more centralized road networks (cf. Fig. 12). The monomial term of betweenness centrality further exhibits a positive correlation with circuity, confirming our intuition that centralized road network designs will lead to less efficient local travel.

On the other hand, node connectivity and circuity are negatively correlated with circuity: the more connected the network, the higher the accessibility to roads, which eventually reduces the need for detours. Nevertheless, for segments with medium levels of

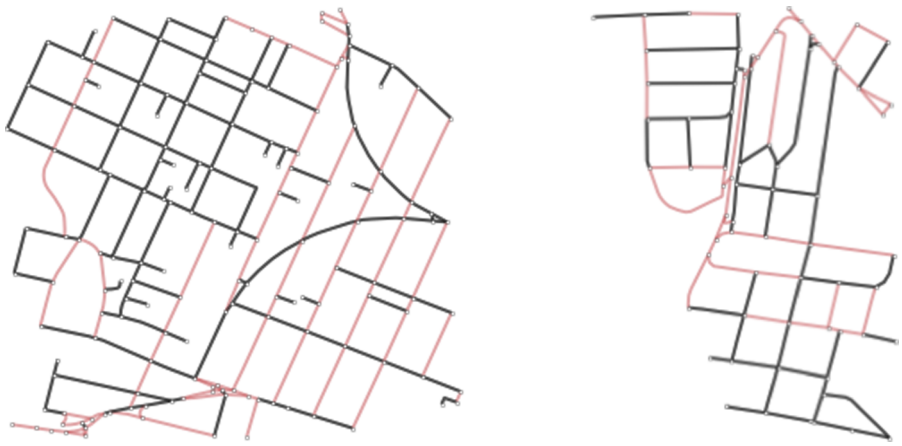


Fig. 12. Sample of road network segments in São Paulo with similar fraction of one-way streets, 40% (red links), and significantly different circuity: 1.79 (left segment) and 3.82 (right segment). This difference in circuity is explained by the fact that betweenness centrality for the right segment is twice as high as it is for the left segment. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

node connectivity, the interaction between this variable and betweenness centrality will tend to increase circuity.

The average node degree also exhibits a decreasing correlation with network circuity. Higher node degree measurements usually indicate closeness to a regular lattice form, and are consequently more efficient compared to tree-like road networks characterized by lower node degree.

The quadratic terms of highway length and node connectivity are statistically significant as well for a significance level of 0.01. The corresponding coefficient signs indicate the concave and convex nature, respectively, of the non-linear relationship with circuity. The significance of the polynomial terms of these variables also emphasizes the importance of both variables in explaining local network circuity. In Appendix A, we validate the results of the polynomial regression by comparing them with the results of a random forest (RF) regression model.

In Table 7 we also report, for each monomial term, the resulting variation Δ in R^2 if the corresponding variable would be excluded form the model. Our goal is to quantify the individual contribution of each variable in explaining the variability in the data. Results indicate that excluding node connectivity from the regression model leads to an 11 percent point (p.p.) reduction in R^2, the largest reduction among all six variables. This observation is consistent with other regression model results as the standardized coefficient with largest magnitude also corresponds to node connectivity, suggesting that this variable is arguably the most relevant in our model. The Δ R^2 corresponding to betweenness centrality is -7 p.p. All other variables result in an R^2 reduction between 2 and 3 percent points. Finally, we verify for multicollinearity in our regression model using two statistical tests: VIF and conditional numbers (see Table 8). None of the VIF for each of the explanatory variables is larger than 10. We also note that none of the conditional numbers is greater than 30, which would have indicated moderate to strong dependencies (Belsley et al., 1980).

5. Generalizing local road network circuity

In this section, we generalize the intricate correlations between the circuity factor c and topological and dimensional properties of the road network observed in São Paulo to other case studies. At a coarse-grained level of analysis, we aim to explore: i) if the difference in circuity between constrained road network segments (i.e., CL1) and fine-grained road network segments (i.e., CL2) holds in other cities, and ii) if the correlation patterns we observe between circuity and road network properties in São Paulo can be generalized to other cities. For this purpose, we collect data for seven additional cities following the same data collection protocols previously described.

We are mindful of the *small-N* nature of our study and, consequently, of the classic criticism on the limitations of case study-based

Table 8
Multicollinearity tests.

Variance Inflation Factor		Conditional Index	
Highway length	1.27	Linear model	3.50
Primary road length	1.30	Polynomial model	23.3
One-way fraction	2.54		
Betweenness centrality	1.08		
Node connectivity	2.83		
Node degree	1.95		
Average VIF	1.83		

research to derive broad generalizations (see Tsang (2014) and references therein). However, as Tsang (2014) notes, case studies are well suited to explore mechanismic explanations. Our approach resonates with his argument.

We focus our analysis on a selected (convenience) sample of cities of different population sizes to generalize the patterns and correlations observed for the São Paulo case. The set of case studies includes urban areas of similar (very large) size, namely Mexico City; three large[2] metropolitan areas: Rio de Janeiro, Lima and Bogotá; and a set of medium-sized cities[3] in Latin America and the US: Quito, Boston and Denver. This selection aims at incorporating different city sizes and geographic contexts in our analysis.

5.1. Generalization based on regression analysis

To analyze if the correlations and significance of variables observed for the São Paulo case are also observed for the other cities, we fit the polynomial regression model of second degree introduced in Section 4.3 with the data corresponding to the other case studies. We exclude Quito in this analysis due to data limitations. Numerical results are presented in Appendix C. Overall, we observe that when the corresponding explanatory variable is significant, the direction of the relationship observed between circuity and the explanatory variable observed for São Paulo also holds true for the other cases. For instance, highway length is significant for $p < 0.05$ in all cases except for Bogotá and is positively correlated with circuity, as previously noted in Section 4.3 for São Paulo. A similar observation is made for betweenness centrality and node connectivity, which are positively and negatively correlated in all cases, respectively, and are also significant in all cases. The correlation direction for node degree, one-way fraction, and for the interaction terms is also consistent with the results observed for São Paulo, yet these terms are significant only for a reduced number of cases. Finally, we also analyze for Denver and Mexico City (case studies with more than 900 city segments) the $\Delta\ R^2$ if each of the variables would be excluded form the model and results are consistent with those observed for São Paulo: node connectivity and betweenness centrality yield the largest p.p. reduction in R^2.

Results based on the regression analysis applied to the additional case studies confirm the direction of the relationship between the dimensional and topological variables and local road network circuity observed for São Paulo. Nonetheless, as noted above, not all explanatory variables were always significant for that particular regression model choice, which prevents us from deriving broader generalizations. In the section below, we propose a methodology based on a classification of urban segments to further explore these correlations at a coarse-grained level of analysis.

5.2. Methodology

We introduce a quantitative method to i) classify urban segments based on road network properties, and ii) conduct comparative analyses. The classification step builds on the generative method for clustering based on GMM introduced in Section 4.2. The comparative analysis step leverages classic statistical methods, namely hypothesis tests on probability distributions and means, to analyze road network circuity (dis) similarities and correlations across case studies.

Classification analysis. We build on the generative GMM introduced in Section 4.2 to classify segments in other urban areas. Specifically, we leverage clusters generated for São Paulo using GMM with $K = 3$ to generate a classifier. We preserve the same unit of geo-spatial analysis, i.e., 1 km^2 segments, and the same set of explanatory variables (see Table 3 and Table 4) used to fit the GMM for São Paulo. The primary goal of this semi-supervised classification method is to generate comparable clusters of urban segments across different cities based upon topological and dimensional road network properties. We refer to this methods as *semi-supervised* as we first use an unsupervised learning model (i.e., GMM) to generate a classifier, and, second, we use this fitted model to *predict* the corresponding class for each segment in the other cities. For validation purposes, we compare these results against those obtained by generating a classifier fitted for each individual city.

Comparative analysis. Once each segment has been classified in one of the $K = 3$ clusters, we use classical statistical methods to conduct intra-city and inter-city comparisons. As in Section 4, we exclude from these analyses segments corresponding to CL3.

First, we explore intra-city differences in circuity by analyzing the conditional probability distribution of c per cluster,

$$f(c) = \sum_{\theta} f(c|\theta)p(\theta), \tag{5}$$

where $\theta = \{\theta_1, \theta_2\}$ for CL1 and CL2, respectively.

For each city j, we conduct a Kolmogorov-Smirnov test (Law and Kelton, 2000) to assess the equality of $f_j(c|\theta_1)$ and $f_j(c|\theta_2)$. The goal is to identify intra-city differences in c between clusters. We define the following null (H_0) and alternative (H_1) hypotheses:

(i) H_0^I: $f_j(c|\theta_1)$ and $f_j(c|\theta_2)$ share the same empirical distribution
(ii) H_1^I: $f_j(c|\theta_1)$ and $f_j(c|\theta_2)$ do not share the same empirical distribution

Furthermore, for inter-city comparisons, we use the mean $\bar{c}$ and variance σ_c^2 to conduct a pair-wise hypothesis test to statistically analyze (dis) similarities in c. Specifically, we conduct a two-sided Welch's t-test for the equality of $\bar{c}$ assuming unequal variances and different population or sample sizes (Law and Kelton, 2000). More formally, for every pair of cities (j, l) and cluster type θ, let $\bar{c}_{j\theta}$ be

[2] cities with at least 9 million inhabitants
[3] cities with 2–5 million inhabitants

the average local circuity factor for city j in cluster θ. Then, we define the null (H_0) and alternative (H_1) hypotheses:

(i) H_0^{II}: the $\bar{c}_{j\theta} = \bar{c}_{l\theta}$, mean circuity factors in cities i, j for cluster θ, are equal
(i) H_1^{II}: the $\bar{c}_{j\theta} \neq \bar{c}_{l\theta}$, mean circuity factors in cities i, j for cluster θ, are not equal

The assumption of unequal variances is verified by means of a Levene test (Law and Kelton, 2000) for equality of variances, using the following null (H_0) and alternative (H_1) hypotheses:

(i) H_0^{III}: the $\sigma^2_{j\theta} = \sigma^2_{l\theta}$, variances of c in cities i, j for cluster θ, are equal
(ii) H_1^{III}: the $\sigma^2_{j\theta} \neq \sigma^2_{l\theta}$, variances of c in cities i, j for cluster θ, are not equal

We use a significance level of $\alpha = 0.10$ for all tests.

5.3. Application

Classification analysis. We apply the classification method described above to all case studies. Fig. 13 shows the spatial distribution of the resulting clusters for each city. In general, spatial distributions of clusters evidence consistency with the results observed for São Paulo: segments corresponding to CL1 (red) cluster inner parts of the city. CL1 also includes segments having a significant fraction of large capacity roads. Segments classified within CL2 correspond to outer city segments where the road network is well connected and less constrained. However, we must be cautious about generalizations for CL2: since this cluster covers 44–53% of the built-up area in these cities, we should expect certain levels of road network heterogeneity among segments even within the same cluster. Finally, as observed in São Paulo, CL3 includes zones in urban edges and other zones with coarse-grained road networks.

To validate the performance of the classification method, we quantify a classification consistency score by comparing the results of the proposed classifier against classification results obtained by fitting the GMM-based clustering method to each case study individually (see Table 9). While a detailed analysis on the classification accuracy of the method falls outside the scope of this study, we argue that our classification method is robust as it yields classification consistency scores between 78% and 95%. We observe higher classification consistency for Mexico City, Rio de Janeiro and Lima, possibly explained by the similarities in city size, geographic location, and socio-economic contexts among these cities. Classifications scores above 0.80 are observed for Bogotá and Quito. While city size might explain lower classification consistency in Quito, differences in build-up area size, and, consequently, population density might explain the score for Bogotá. These differences in build-up area size amplify for Boston and Denver, hence the lower classification consistency scores.

Comparative analysis. In examining the conditional probability distribution depicted in Fig. 14, we make the following observations. In each city, $f(c|\theta_1)$ is shifted to the right compared to $f(c|\theta_2)$, suggesting higher values of circuity for CL1. These differences are statistically verified by means of the KS hypothesis test on the equality of empirical distributions for $f(c|\theta_1)$ and $f(c|\theta_2)$ per city. Our test results reveal that H_0^1 is rejected for all cases ($\alpha = 0.10$), confirming that in all eight cities, segments in CL1 will exhibit significantly higher levels of road network circuity.

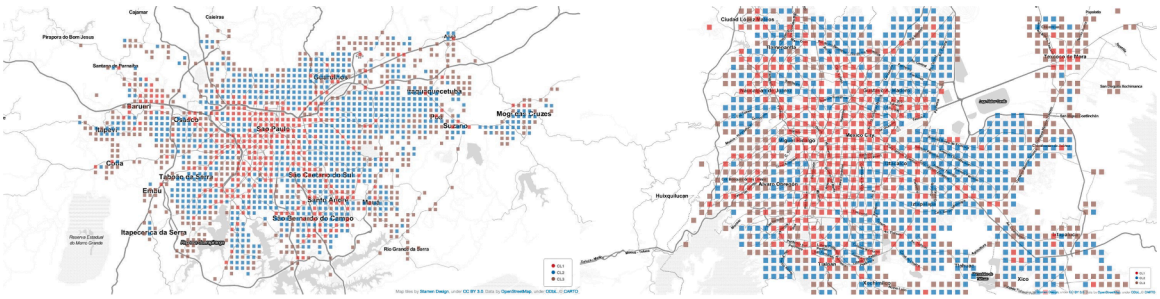
Next, for each cluster, we analyze inter-city differences in circuity, based on the pair-wise hypothesis two-sided Welch's t-tests (see Fig. 15). We complement this analysis by further exploring correlations between c and key topological and dimensional covariates (average values reported in Tables 10 and 11 for clusters CL1 and CL2, respectively).

In a pre-processing step, we conduct Levene tests to assess the unequal variances assumption. For CL1, H_0^{II} is only rejected ($\alpha = 0.10$) for pair-wise comparisons that included the city of Denver. For CL2, H_0^{II} is rejected in most pair-wise comparisons. Thus, we argue that the assumption of unequal variances accounts for the most general case and should be used for the Welch's t-tests for the equality of $\bar{c}$ assuming unequal variances.

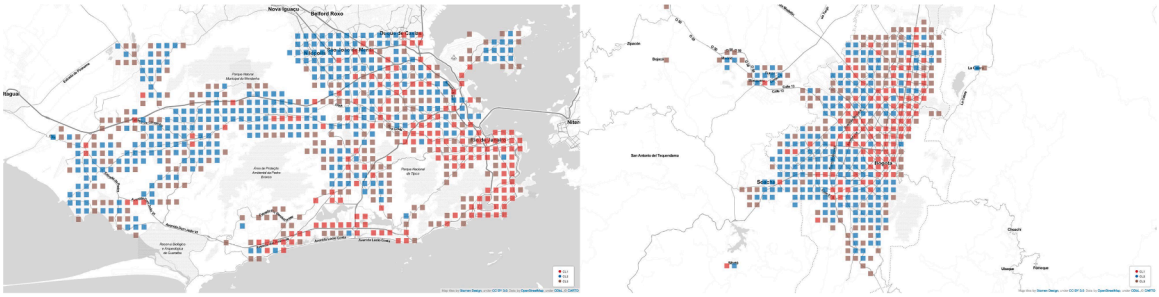
Based on our Welch's t-test results for CL1 (see Fig. 15), H_0^{II} can not be rejected for the subset including São Paulo-Mexico City-Bogotá, and for the subset Quito-Lima-Boston. We reject H_0^{II} for all pair-wise comparisons for the cities of Rio de Janeiro and Denver, which is not surprising, given that these two cities exhibit the highest and lowest $\bar{c}$, respectively (see Table 10).

Multiple factors explain the low average circuity for Denver: it exhibits the lowest values for one-way fraction and primary-road length. Most importantly, Denver exhibits the highest node connectivity of all case studies (possibly because it is the youngest of all cities analyzed). We argue that the combination of these factors drives relatively lower circuity levels in CL1 in Denver. On the contrary, Rio de Janeiro exhibits the largest average circuity. In Table 10, we observe that Rio's segments in CL1 exhibit the lowest levels of network connectedness both in terms of node degree and node connectivity. These two contrasting examples evidence the impact of the connectivity of the network on circuity. When comparing the subsets {São Paulo, Mexico City, Bogotá} against {Lima, Quito, Boston}, differences between these two groups are driven by highway length and node connectivity. As expected, larger highway road length and lower average node connectivity will increase the mean circuity for the {São Paulo, Mexico City, Bogotá} subset.

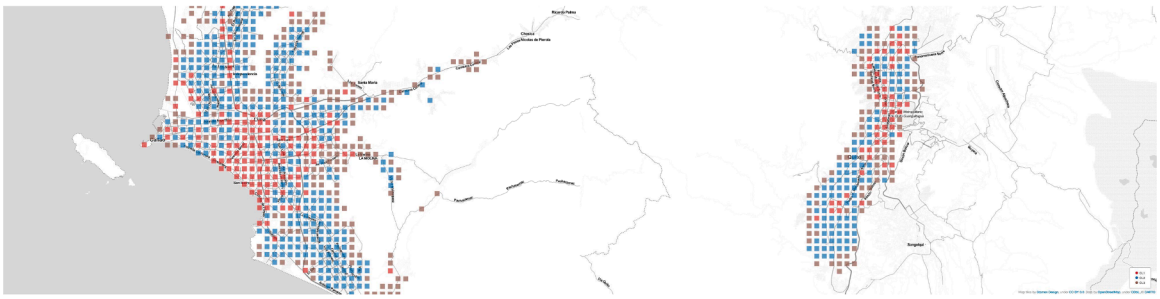
For cluster CL2, Welch's t-test indicates four pairs of cities for which H_0^{II} is not rejected: {Rio de Janeiro, Bogotá}; {São Paulo, Lima}; {Quito, Mexico City}; and {Boston, Denver} (see Fig. 16). The pair {Rio de Janeiro, Bogotá} exhibits the highest $\bar{c}$ (see Table 11). While higher primary road length plausibly explains higher circuity in Bogotá, lower node degree and node connectivity values explain higher circuity in Rio de Janeiro. Similarly, for the pair {Boston, Denver}, which exhibits the lowest $\bar{c}$, lower highway



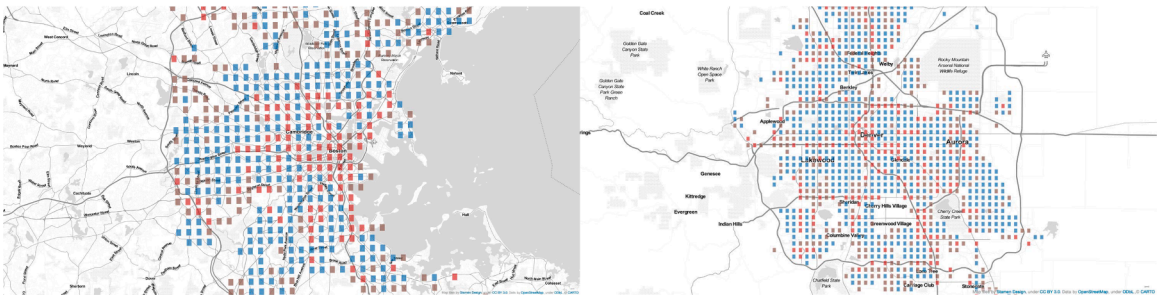
(a) Sao Paulo (b) Mexico City



(c) Rio de Janeiro (d) Bogotá



(e) Lima (f) Quito



(g) Boston (h) Denver

Fig. 13. Spatial distribution of clusters CL1 (red), CL2 (blue) and CL3 (brown) across case studies. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 9
Classification consistency scores.

	City							
	São Paulo	Mexico City	Rio de Janeiro	Lima	Bogotá	Quito	Boston	Denver
Score	1	0.94	0.94	0.95	0.81	0.85	0.78	0.79

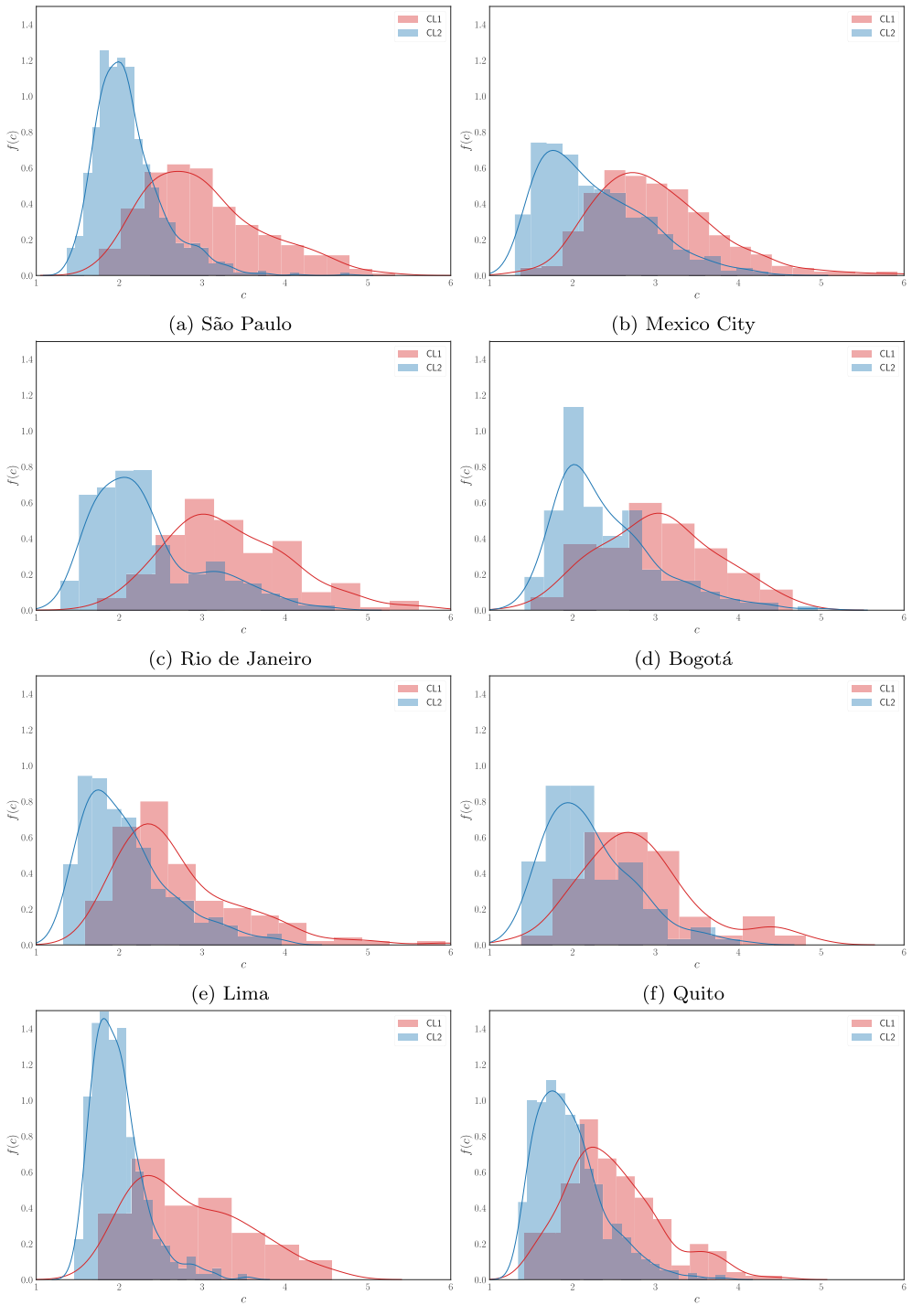


Fig. 14. Conditional probability distributions of local circuity for cities.

length for Denver and lower primary road length for Boston explain the circuity levels observed. These results confirm the general correlation patterns concluded in Section 4 and also confirm the intricate correlation between road network properties and circuity. Future research should explore the magnitude of individual and/or combined effects of these different variables on road network circuity across case studies.

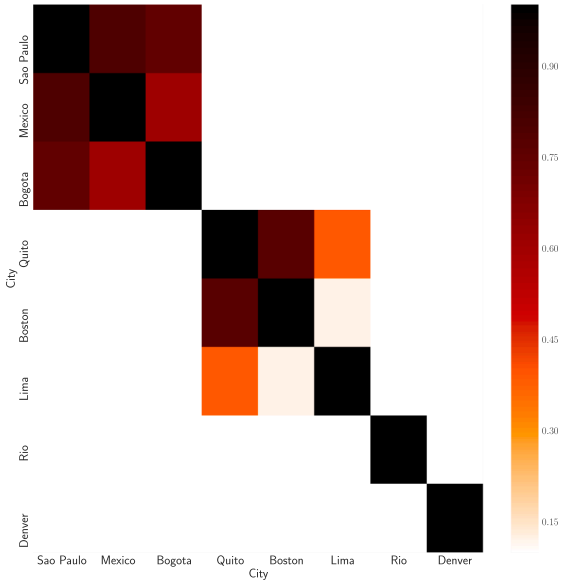


Fig. 15. *p*-value heat-map of inter-city Welch's *t*-tests for cluster CL1. Colored cells indicate pair-wise test for which we do not reject the H_o ($\alpha = 0.10$).

Table 10
Average values of segment-level variables for cluster CL1.

	São Paulo	Mexico City	Rio	Lima	Bogotá	Quito	Boston	Denver
Circuity factor $\bar{c}$	3.02	3.00	3.33	2.70	3.04	2.81	2.84	2.49
One-way fraction (%)	0.43	0.49	0.50	0.54	0.50	0.39	0.42	0.34
Highway length (km)	964	896	920	718	916	518	755	828
Primary road length (km)	1,518	1,375	1,009	1,457	2,082	2,372	1,586	733
Node degree	3.10	3.11	3.00	3.25	3.16	3.03	3.04	3.11
Node connectivity	0.97	0.94	0.83	1.10	0.98	1.05	0.98	1.14
Betweenness centrality	0.07	0.07	0.08	0.07	0.06	0.07	0.09	0.10
Fraction of urban area	0.20	0.21	0.23	0.19	0.24	0.21	0.17	0.17
Amb. population (inh/km^2)	15,005	14,510	10,059	17,774	25,195	12,350	6,625	2,562

Table 11
Average values of segment-level variables for cluster CL2.

	São Paulo	Mexico City	Rio	Lima	Bogotá	Quito	Boston	Denver
Circuity factor $\bar{c}$	2.12	2.25	2.33	2.09	2.41	2.20	1.98	1.97
One-way fraction (%)	0.10	0.10	0.08	0.18	0.15	0.14	0.15	0.07
Highway length (km)	90	112	78	60	166	54	165	12
Primary road length (km)	361	309	274	472	599	566	114	205
Node degree	3.03	3.03	2.93	3.16	3.01	3.00	2.90	2.98
Node connectivity	1.56	1.58	1.47	1.71	1.51	1.54	1.41	1.57
Betweenness centrality	0.06	0.06	0.07	0.05	0.05	0.06	0.08	0.09
Fraction of urban area	0.48	0.45	0.47	0.48	0.46	0.44	0.46	0.53
Amb. population (inh/km^2)	11,550	12,960	11,550	12,740	21,130	6,500	3,640	1,960

6. Conclusion

At the local level, the efficiency of the road network is explained by several dimensional and topological properties, some of which vary considerably across a city. Local circuity factors capture these complex interactions in a simple measure, which can be used to improve shortest path distance approximations, but also to better understand how the topological and physical properties of the street network impact travel directness, and, consequently inform logistics practice and urban transportation policy.

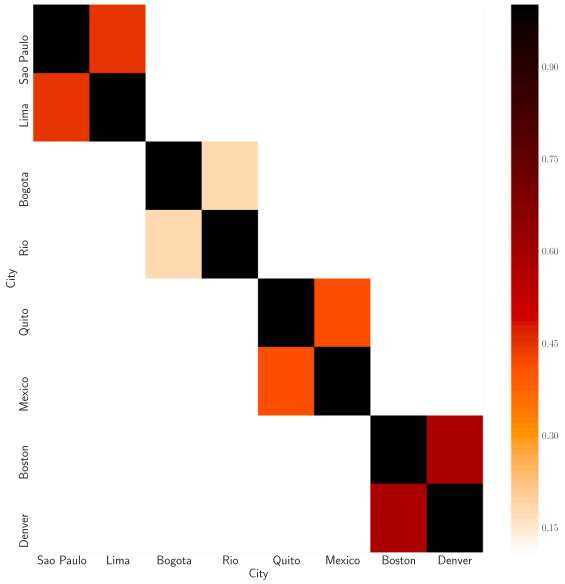


Fig. 16. *p*-value heat-map of inter-city Welch's *t*-tests for cluster CL2. Colored cells indicate pair-wise test for which we do not reject the H_o ($\alpha = 0.10$).

Leveraging the metropolitan area of São Paulo, Brazil, as the primary example, we observe a significant heterogeneity of road network circuity across the city. Using 1-km^2 segments as the unit of geo-spatial analysis and a large sample of real shortest-path trips extracted from the Google Distance Matrix service, we derive values for circuity (c) that range between 1.38 and 5.32, with $\bar{c} = 2.51$. The magnitude and range of these results unveil two important insights. First, on average, real trip distances are about twice as long as distances predicted by the L_1 norm, suggesting that the assumptions encoded in this norm ($\bar{c} = 1.27$) significantly oversimplify the underlying real road-network. Second, a single city-wide measurement of circuity (cf. Table 1) fails to capture the heterogeneity in travel efficiency observed across the city. While a city-wide circuity measurements might provide a good approximation for the 'line-haul' portion of a route (which resembles commuter travel patterns), these same measurements would not yield robust distance estimates for the 'local delivery and pickup' portion of the route.

The explanatory regression model introduced in Section 4 derives correlations between circuity and dimensional and topological properties of the road network. Large-capacity roads (highways and primary roads) exhibit a positive correlation with local circuity. In contrast to city-wide trips, in which large capacity roads facilitate travel directness, locally these types of roads complicate travel due to their reduced accessibility. However, the efficiency of the road network for local trips is not only driven by obstacles. Other complications to travel directness, such as one way streets, and the topology itself of the road network also impact travel efficiency. On the other hand, a better connected street network, measured by its average node connectivity and node degree, increase street accessibility and, therefore, leads to more efficient travel. As discussed in Section 5, these correlations between road-network topological and dimensional properties and local circuity are consistent, at different levels of magnitude, across a selected set of additional cities in Latin America and the US analyzed in this paper.

In Section 4 we introduce a classification of urban segments according to road network properties. Three categories are proposed: constrained, fine-grained and coarse-grained, corresponding to approximately 20%, 50% and 30% of the urban area respectively. Constrained areas should be given special attention in designing last mile distribution systems and in overall traffic management: constrained zones generally exhibit higher levels of local circuity and higher levels of population density, which implies that a disproportionate portion of urban logistics flows concentrates in a fraction of urban areas with lower road network efficiency.

New large traffic and road network data sets such as the Google Distance Matrix service and OpenStreetMaps are opening new frontiers for large-scale quantitative analysis of urban problems. Still, data completeness and quality need to be verified, particularly if datasets have been collected through collaborative, open-licensed initiatives as with OSM. In our primary case study São Paulo, only minor inconsistencies in the road network dataset were found. However, the reliability of such data might vary from one city to another.

Finally, while this paper has been inspired by the network design challenges faced by e-retailers and manufacturers serving urban customers and consumers through last-mile delivery networks, insights derived from this research are transferable to any route-based urban transportation systems serving a large customer base. Examples of such services include school bus systems and, more recently, ride-sharing systems. Thus, we argue that the relevance of studying the local efficiency of urban road networks spans multiple

transportation applications and entails relevant implications for urban transportation/logistics practice and policy. A better understanding of the efficiency of local trips can inform, for instance, logistics service strategies, traffic management interventions, or road network design choices.

CRediT authorship contribution statement

Daniel Merchán: Conceptualization, Methodology, Validation, Investigation, Formal analysis, Writing - original draft, Writing - review & editing, Visualization. **Matthias Winkenbach:** Conceptualization, Methodology, Formal analysis, Writing - review & editing. **André Snoeck:** Methodology, Software, Data curation, Visualization, Writing - review & editing.

Acknowledgment

The authors thank Prof. Jan C. Fransoo and two anonymous reviewers for their constructive comments. The authors also thank Dr. Edgar E. Blanco for his valuable feedback in early stages of this work.

Appendix A. Random forest regression

We further validate our results from the regression model presented in Section 4.3 by comparing it against a RF regression (Breiman, 2001). Even though random forests are better suited for predictive rather than explanatory models, they offer two benefits to our circuity analysis: 1) a ranking of relative importance of the explanatory variables for prediction purposes, and 2) a benchmark regression model that does not enforce any mathematical form. We fit the RF regression model using the Scikit-learn Python module (Pedregosa et al., 2012). The RF model yields $R^2 = 0.86$ for train and test sets (number of trees $= 80$, depth $= 8$).

Node connectivity is the single most relevant predictor in the RF model (Fig. A.17). This result is consistent with the quadratic regression model (cf. Table 7) in which node connectivity is the variable with the largest coefficient in magnitude, followed by betweenness centrality. Interestingly, the dimensional variables have relatively lower importance in the RF model. This is explained by the 'clumped-at-zero' nature of dimensional variables (see Fig. 11). That is to say, while their correlation with circuity is significant, there is a large number of segments in which, for instance, the value for highway length is zero.

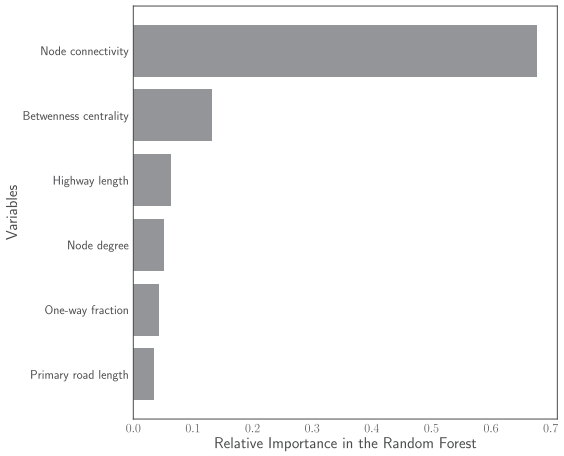


Fig. A.17. Ranking of importance of the explanatory variables in the RF regression.

Appendix B. Formulae

In this section we provide the formulae corresponding to the topological variables listed in Table 4. For all variable definitions, let $n \in N$ be a node in the network corresponding to city segment $i \in I$ where $|N|$ is the cardinality of set N.

Node connectivity. Let γ_{st} be the number of nodes to remove to disconnect two non-adjacent nodes s and t. γ_{st} is obtained using a maximum flow algorithm on an auxiliary digraph build from the original graph (Kammer et al., 2005). The node connectivity γ_i for city segment i is then obtained by averaging γ_{st} for all pairs of non-adjacent nodes in the graph.

Node degree. Let κ_n be degree of node n, i.e., the number of edges emanating from it. The average node degree for city segment i is given by

$$\kappa_i = \frac{\sum_n \kappa_n}{|N|}. \tag{B.1}$$

Neighborhood Degree. Let $S \in N$ be the subset of nodes connected to node n. The neighborhood degree η_n is defined as

$$\eta_n = \frac{\sum_s \kappa_s}{|S|}, \tag{B.2}$$

and the average neighborhood degree for city segment i is then given by

$$\eta_i = \frac{\sum_n \eta_n}{|N|}. \tag{B.3}$$

Betweenness centrality. The betweenness centrality g_n for node n is defined as

$$g_n = \sum_{s \neq t} \frac{\theta_{st}(n)}{\theta_{st}}, \tag{B.4}$$

where θ_{st} is the number of shortest paths from s to t and $\theta_{st}(n)$ is the number of shortest paths from s to t through node n. Then the average betweenness centrality g_i for city segment i is given by

$$g_i = \sum_n \frac{g_n}{|N|}. \tag{B.5}$$

Closeness centrality. Let d_{ns} be the length of the shortest path between nodes n and s. Thus the average closeness centrality m_i for city segment i is defined as

$$m_i = \frac{\sum_n [\sum_s d_{ns}]^{-1}}{|N|} \tag{B.6}$$

Degree centrality. Let l_n be the fraction of nodes in N that node n is connected to. Then the average degree centrality of the network corresponding to city segment i is defined as

$$l_i = \frac{\sum_n l_n}{|N|}. \tag{B.7}$$

Appendix C. Regression results

Tables C.12–C.17.

Table C.12
Results form the regression model with standardized values for Mexico City.

	Coeff. (β)	Std. Err.	p-value	ΔR^2 (p.p.)
Intercept	2.479	0.033	0.000	**
X_1 Highway length	0.148	0.035	0.000	−2
X_2 Primary road length	0.043	0.034	0.204	-1
X_3 One-way fraction	0.203	0.042	0.000	−3
X_4 Betweenness centrality	0.147	0.023	0.000	-3
X_5 Node connectivity	−0.326	0.040	0.000	−8
X_6 Node degree	−0.122	0.031	0.000	−2
X_1^2 Highway length × Highway length	−0.007	0.009	0.433	**
X_5^2 Node connectivity × Node connectivity	0.147	0.043	0.001	**
X_3X_4 One-way fraction × Betweenness centrality	0.095	0.023	0.000	**
R^2: 0.50	F-statistic:	41.66		
Adj. R^2: 0.48	Obs.: 1170			

Table C.13
Results form the regression model with standardized values for Rio de Janeiro.

	Coeff. (β)	Std. Err.	p-value
Intercept	2.567	0.054	0.000
X_1 Highway length	0.361	0.055	0.000
X_2 Primary road length	0.065	0.051	0.204
X_3 One-way fraction	0.096	0.073	0.189
X_4 Betweenness centrality	0.132	0.034	0.000
X_5 Node connectivity	−0.297	0.066	0.000
X_6 Node degree	−0.095	0.044	0.031
X_1^2 Highway length × Highway length	−0.004	0.014	0.776
X_5^2 Node connectivity × Node connectivity	0.010	0.059	0.094
X_3X_4 One-way fraction × Betweenness centrality	0.031	0.052	0.544
R^2: 0.57	F-statistic: 24.12		
Adj. R^2: 0.55	Obs.: 512		

Table C.14
Results form the regression model with standardized values for Lima.

	Coeff. (β)	Std. Err.	p-value
Intercept	2.268	0.041	0.000
X_1 Highway length	0.201	0.051	0.000
X_2 Primary road length	0.075	0.041	0.070
X_3 One-way fraction	0.002	0.053	0.962
X_4 Betweenness centrality	0.298	0.049	0.000
X_5 Node connectivity	- 0.217	0.049	0.000
X_6 Node degree	−0.047	0.043	0.281
X_1^2 Highway length × Highway length	−0.004	0.014	0.751
X_5^2 Node connectivity × Node connectivity	0.068	0.053	0.198
X_3X_4 One-way fraction × Betweenness centrality	0.043	0.047	0.355
R^2: 0.47	F-statistic: 15.95		
Adj. R^2: 0.44	Obs.: 519		

Table C.15
Results form the regression model with standardized values for Bogotá.

	Coeff. (β)	Std. Err.	p-value
Intercept	2.396	0.058	0.000
X_1 Highway length	0.031	0.065	0.640
X_2 Primary road length	0.023	0.053	0.658
X_3 One-way fraction	0.111	0.071	0.119
X_4 Betweenness centrality	0.273	0.037	0.000
X_5 Node connectivity	−0.426	0.09	0.000
X_6 Node degree	0.044	0.051	0.389
X_1^2 Highway length × Highway length	0.007	0.032	0.825
X_5^2 Node connectivity × Node connectivity	0.262	0.061	0.000
X_3X_4 One-way fraction × Betweenness centrality	0.090	0.053	0.089
R^2: 0.53	F-statistic: 12.03		
Adj. R^2: 0.49	Obs.: 317		

Table C.16
Results form the regression model with standardized values for Boston.

	Coeff. (β)	Std. Err.	p-value
Intercept	2.068	0.032	0.000
X_1 Highway length	0.211	0.045	0.000
X_2 Primary road length	−0.041	0.035	0.233
X_3 One-way fraction	0.071	0.035	0.044
X_4 Betweenness centrality	0.160	0.030	0.000
X_5 Node connectivity	−0.214	0.033	0.000
X_6 Node degree	−0.014	0.028	0.613
X_1^2 Highway length × Highway length	−0.055	0.017	0.001
X_5^2 Node connectivity × Node connectivity	0.064	0.029	0.028
X_3X_4 One-way fraction × Betweenness centrality	0.04	0.028	0.161
R^2: 0.70	F-statistic: 32.78		
Adj. R^2: 0.67	Obs.: 416		

Table C.17
Results form the regression model with standardized values for Denver.

	Coeff. (β)	Std. Err.	p-value	Δ R^2 (p.p.)
Intercept	2.040	0.020	0.000	**
X_1 Highway length	0.150	0.031	0.000	−5
X_2 Primary road length	0.008	0.019	0.685	−1
X_3 One-way fraction	0.123	0.023	0.000	−4
X_4 Betweenness centrality	0.133	0.016	0.000	−5
X_5 Node connectivity	−0.152	0.028	0.000	−6
X_6 Node degree	−0.132	0.024	0.000	−3
X_1^2 Highway length × Highway length	−0.005	0.007	0.462	**
X_5^2 Node connectivity × Node connectivity	0.192	0.0037	0.000	**
X_3X_4 One-way fraction × Betweenness centrality	0.044	0.013	0.001	**
R^2: 0.61	F-statistic:	52.25		
Adj. R^2: 0.60	Obs.: 941			

References

Ansari, S., Başdere, M., Li, X., Ouyang, Y., Smilowitz, K., 2018. Advancements in continuous approximation models for logistics and transportation systems:1996–2016. Transp. Res. Part B: Methodol. 107, 229–252.
Ballou, R.H., Rahardja, H., Sakai, N., 2002. Selected country circuity factors for road travel distance estimation. Transp. Res. Part A: Policy Pract. 36 (9), 843–848.
Barabási, A.L., 2016. Network Science. Cambridge University Press, Cambridge.
Barthélemy, M., 2011. Spatial networks. Phys. Rep. 499 (1–3), 1–101.
Barthélemy, M., Flammini, A., 2008. Modeling urban street patterns. Phys. Rev. Lett. 100 (13), 1–4.
Bast, H., Carlsson, E., Eigenwillig, A., Geisberger, R., Harrelson, C., Raychev, V., Viger, F., 2010. Fast routing in very large public transportation networks using transfer patterns. In: Proceedings of the 18th Annual European Symposium on Algorithms. vol. 6346. pp. 290–301.
Bast, H., Delling, D., Goldberg, A., Müller-Hannemann, M., Pajor, T., Sanders, P., Wagner, D., Werneck, R.F., 2016. In: Kliemann, L., Sanders, P. (Eds.), 9220. Springer, Cham, Berlin, pp. 19–80.
Batty, M., 2013. The New Science of Cities. MIT Press, Cambridge, MA.
Beardwood, J., Halton, J.H., Hammersley, J.M., 1959. The shortest path through many points. Math. Proc. Cambridge Philos. Soc. 55 (4), 299–328.
Belsley, D., Kuh, E., Welsch, R.E., 1980. Regression Diagnosis: Identifying Influential Data and Sources of Collinearity. John Wiley & Sons, Hoboken, NJ.
Bergmann, F.M., Wagner, S.M., Winkenbach, M., 2020. Integrating first-mile pickup and last-mile delivery on shared vehicle routes for efficient urban e-commerce distribution. Transp. Res. Part B: Methodol. 131, 26–62.
Blanco, E.E., Fransoo, J.C., 2017. Supplying nanostores. In: Fransoo, J.C., Blanco, E.E., Mejia Argueta, C. (Eds.), Reaching 50 Million Nanostores: Retail Distribution in Emerging Megacities. CreateSpace Independent Publishing Platform, pp. 19–38.
Boeing, G., 2017. OSMnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks. Comput. Environ. Urban Syst. 65, 126–139.
Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32.
Bright, E.A., Coleman, P.R., Rose, A.N., Urban, M.L. LandScan. 2015. http://www.ornl.gov/landscan/.
Brimberg, J., Love, R.F., 1992. A New Distance Function for Modeling Travel Distances in a Transportation Network. Transp. Sci. 26, 129–138.
Chien, T.W., 1992. Operational estimators for the length of a traveling salesman tour. Comput. Oper. Res. 19 (6), 469–478.
Crucitti, P., Latora, V., Porta, S., 2006. Centrality measures in spatial networks of urban streets. Phys. Rev. E Stat. Nonlinear Soft Matter Phys. 73 (3), 1–5.
Dablanc, L., 2007. Goods transport in large European cities: difficult to organize, difficult to modernize. Transp. Res. Part A: Policy Pract. 41 (3), 280–285.
Daganzo, C.F., 1984a. The distance traveled to visit N points with a maximum of C stops per vehicle: An analytic model and an application. Transp. Sci. 18 (4), 331–350.
Daganzo, C.F., 1984b. The length of tours in zones of different shapes. Transp. Res. Part B: Methodol. 18 (2), 135–145.
Dijkstra, E.W., 1959. A Note on Two Problems in Connexion with Graphs. Numer. Math. 1 (1), 269–271.
Ehmke, J.F., Campbell, A.M., 2014. Customer acceptance mechanisms for home deliveries in metropolitan areas. Eur. J. Oper. Res. 233 (1), 193–207.
Figliozzi, M., 2008. Planning approximations to the average length of vehicle routing problems with varying customer demands and routing constraints. Transp. Res. Rec. J. Transp. Res. Board 2089, 1–8.
Figliozzi, M., 2009. Planning approximations to the average length of vehicle routing problems with time window constraints. Transp. Res. Part B: Methodol. 43 (4),

438–447.
Fransoo, J.C., Blanco, E.E., 2017. 50 million nanostores. In: Fransoo, J.C., Blanco, E.E., Mejía Argueta, C. (Eds.), Reaching 50 Million Nanostores: Retail Distribution in Emerging Megacities. CreateSpace Independent Publishing Platform, pp. 3–17.
Giacomin, D.J., Levinson, D.M., 2015. Road network circuity in metropolitan areas. Environ. Plan. B: Plan. Des. 42 (6), 1040–1053.
Google, Google Distance Matrix API. 2017. https://developers.google.com/maps/documentation/distance-matrix/.
Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning, second ed. Springer, New York.
Hillier, B., Hanson, J., 1984. The Social Logic of Space. Cambridge University Press, Cambridge.
Huang, J., Levinson, D., 2015. Circuity in urban transit networks. J. Transp. Geogr. 48, 145–153.
Jain, A.K., 2010. Data clustering: 50 years beyond K-means. Pattern Recogn. Lett. 31 (8), 651–666.
Janjevic, M., Winkenbach, M., 2020. Characterizing urban last-mile distribution strategies in mature and emerging e-commerce markets. Transp. Res. Part A: Policy Pract. 133, 164–196.
Janjevic, M., Winkenbach, M., Merchán, D., 2019. Integrating collection-and-delivery points in the strategic design of urban last-mile e-commerce distribution networks. Transp. Res. Part E: Logist. Transp. Rev. 131, 37–67.
Jiang, B., Claramunt, C., 2004. Topological analysis of urban street networks. Environ. Plan. B: Plan. Des. 31 (1), 151–162.
Joerss, M., Schröder, J., Neuhaus, F., Klink, C., Mann, F., 2016. Parcel delivery: The future of last mile. Technical Report September. McKinsey&Company.
Kammer, F., Täubig, H., 2005. Connectivity. In: Brandes, U., Erlebach, T. (Eds.), Network Analysis. Lecture Notes in Computer Science, vol. 3418. Springer, Berlin, pp. 143–177.
Kwon, O., Golden, B., Wasil, E., 1995. Estimating the length of the optimal TSP tour: an empirical study using regression and neural networks. Comput. Oper. Res. 22 (10), 1039–1046.
Lämmer, S., Gehlsen, B., Helbing, D., 2006. Scaling laws in the spatial structure of urban road networks. Physica A 363 (1), 89–95.
Larson, R.C., Odoni, A., 1981. Urban Operations Research, first ed. Prentice Hall, Upper Saddle River, New Jersey.
Law, A.M., Kelton, D., 2000. Simulation Modeling and Analysis, third ed. McGraw-Hill, New York.
Levinson, D., El-Geneidy, A., 2009. The minimum circuity frontier and the journey to work. Reg. Sci. Urban Econ. 39 (6), 732–738.
Louf, R., Barthelemy, M., 2014. A Typology of street patterns. J. R. Soc. Interface 11 (20140924).
Love, R.F., Morris, J.G., 1972. Modelling inter-city road distances by mathematical functions. Oper. Res. Quart. 23 (1), 61–71.
Love, R.F., Morris, J.G., 1979. Mathematical models of road travel distances. Manage. Sci. 25 (2), 130–139.
Merchán, D., Winkenbach, M., 2019. An empirical validation and data-driven extension of continuum approximation approaches for urban route distances. Networks 73 (February), 418–433.
Newell, G.F., 1980. Traffic Flow on Transportation Networks. The MIT Press, Cambridge, MA.
Newell, G.F., Daganzo, C.F., 1986a. Design of multiple vehicle delivery tours - II: Other metrics. Transp. Res. Part B: Methodol. 20 (5), 345–363.
Newell, G.F., Daganzo, C.F., 1986b. Design of multiple-vehicle delivery tours-I a ring-radial network. Transp. Res. Part B: Methodol. 20 (5), 345–363.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, d., 2012. Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. 12, 2825–2830.
Perktold, J., Seabold, S., Taylor, J., 2017. StatsModels: Statistics in Python.
Porta, S., Crucitti, P., Latora, V., 2006. The network analysis of urban streets: A dual approach. Physica A 369 (2), 853–866.
Porta, S., Crucitti, P., Latora, V., 2006. The network analysis of urban streets: A primal approach. Environ. Plan. B: Urban Anal. City Sci. 33 (5), 705–725.
Porta, S., Strano, E., Iacoviello, V., Messora, R., Latora, V., Cardillo, A., Wang, F., Scellato, S., 2009. Street centrality and densities of retail and services in Bologna, Italy. Environ. Plan. B: Plan. Des. 36 (3), 450–465.
Ratti, C., 2004. Space syntax: Some inconsistencies. Environ. Plan. B: Plan. Des. 31 (4), 487–499.
Singleton, A.D., Spielman, S.E., Folch, D.C., 2018. Urban Analytics. SAGE, Los Angeles.
Smilowitz, K., Daganzo, C.F., 2007. Continuum approximation techniques for the design of integrated package distribution systems. Networks 50 (3), 183–196.
The OpenStreetMap Foundation, 2017. OpenStreetMap. http://www.openstreetmap.org/.
Tsang, E.W., 2014. Generalizing from research findings: The merits of case studies. Int. J. Manage. Rev. 16 (4), 369–383.
United States Department of Transportation, 2009. National Household Travel Survey. Technical Report.
Winkenbach, M., Janjevic, M., 2018. Classification of last-mile delivery models for e-commerce distribution - a global perspective. In: Taniguchi, E., Thompson, R.G. (Eds.), City Logistics 1: New Opportunities and Challenges. Wiley-ISTE, pp. 162–176.
Winkenbach, M., Kleindorfer, P.R., Spinler, S., 2016a. Enabling Urban Logistics Services at La Poste through Multi-Echelon Location-Routing. Transp. Sci. 50 (2), 520–540.
Winkenbach, M., Roset, A., Spinler, S., 2016b. Strategic redesign of urban mail and parcel networks at La Poste. Interfaces 46 (5), 445–458.